The role of hydrophobicity in protein folding and aggregation

Dijk, E.

2017

Vrije Universiteit

The Role of Hydrophobicity in Protein Folding and Aggregation

Academisch proefschrift

Ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof. dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op woensdag 20 september 11:45
in de aula van de Universiteit,
De Boelelaan 1105

door

Erik van Dijk

Geboren te Heemstede

Promotor: prof. dr. Jaap Heringa
Co-promotor: dr. Sanne Abeln

Contents

1 Introduction **1**

2 The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures **9**
2.1 Supplemental information -The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures 25

3 Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation **27**
3.1 Supplementary Material – Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation . . 38

4 BICEP: Heat Capacity Baseline Prediction **53**
4.1 Supporting Information - BICEP: Heat Capacity Baseline Prediction 65

5 Predicting the hydrophobic surface area of native protein structures from sequence **71**
5.1 Supplementary Information - Predicting the hydrophobic surface area of native protein structures from sequence 86

6 Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence **91**
6.1 Supplemental information - Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence . . 104

7 Discussion **107**

Bibliography **113**

Chapter 1

Introduction

The central dogma in biology is 'DNA makes RNA makes protein'. In simple terms, DNA holds the blueprint to a protein, RNA makes a copy of this blueprint and transfers it to a protein "factory" called a ribosome, and the proteins perform a function. This mechanism is present in not just humans and animals, but in all organisms, including bacteria and plants. Proteins are involved in the majority of processes in all organisms. Some examples: oxygen transport (myoglobin), regulation of the blood sugar level (insulin), and the digestion of proteins present in our food (pepsin). To perform these various functions, proteins need to fold into a specific 3-dimensional shape, which is often referred to as a protein's "native" or "folded" structure. On the surface of this native structure, there is typically some region that performs a protein's function. In our examples, myoglobin has an active region where it captures an oxygen molecule, insulin contains a region that breaks down sugar, and pepsin has a region that interacts a-specifically with other proteins, and breaks them into small fragments, allowing them to recycled in our body to make new proteins (Kendrew et al., 1958; Blundell et al., 1971; Cooper et al., 1990).

For a protein, obtaining the right fold is crucial to it's function. This is illustrated by many diseases that are caused by a single misfolded protein. One of them, cystic fibrosis is primarily caused by a single hereditary point mutation, ΔF508. The mutation prevents the anchoring of the protein to the cell membrane, impeding its function in the cell. This single amino acid therefore prevents the protein from performing its function.

Loss of function is not the only way in which protein misfolding is known to cause disease. Sometimes, for reasons that are not entirely clear, proteins can clump together. This process is called aggregation. For example, in Parkinson's disease, α-synuclein aggregates are considered to cause the brain damage that manifests itself in patients suffering from this disease.

The aim of this thesis is to understand the physical mechanisms that govern protein folding and aggregation. Like everything in the universe, the proteins in our body are subject to the laws of nature. If we can discover the physical principles that govern protein folding, this should allow us to better understand what conditions cause a protein to fold, to fold incorrectly, or to aggregate. Predicting what fold a protein will obtain from its composition is called the 'protein folding problem'. Solving this problem is sometimes considered to be

the 'holy grail' in computational biology. To be able to contribute something to the large and expanding literature, we narrowed our focus to investigate the effect of hydrophobic amino acids, and the role they play in the stability of proteins at different temperatures.

We used approaches from diverse fields to study the contribution of the hydrophobic effect in protein folding and aggregation: physics (Chapter 2), data mining (Chapter 3), statistics (Chapter 4), machine learning (Chapter 5) and simulations (Chapter 2 and 6). Using the right technique to study a phenomenon helps present us a more complete view of the role the temperature dependence of the hydrophobic effect in protein folding and aggregation.

To understand the chapters in this thesis, it is necessary to understand the chemical composition of proteins, as well as the forces that produce their native structure.

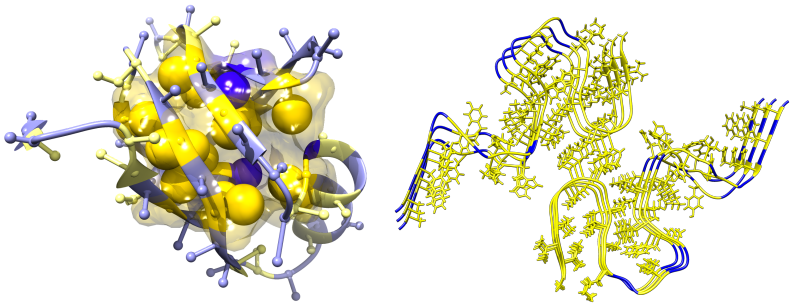

FIGURE 1.1: **Aggregated state and folded state**. Two alternative states for proteins, the functional native state of a protein with PDB-ID:2K5I (left) and the toxic disease causing amyloid fibre aβ-42 (PDB-ID: 2NAO) (Wälti et al., 2016) (right). Hydrophobic residues are colored in yellow, while hydrophilic residues are colored in blue. In a typical protein, a hydrophobic core can be observed, where the hydrophobic residues are shielded from the water by the hydrophilic residues. Aggregating proteins, and especially the core aggregating region of proteins tend to be more hydrophobic than normal proteins. The lack of protection from hydrophilic amino acids can cause a runaway aggregation of proteins (see left side).

From amino acids to proteins

Amino acids consist of a backbone consisting a nitrogen atom followed by two carbon atoms. ("Amino-and-then-acid"). One carbon atom is connected to an oxygen, and the other is connected to a side chain. The side chain is what makes each of the 20 amino acids unique. In a typical protein, the chains are connected to each other by forming a bond that removes a water molecule from

the amino acids and connects the C_β-atom from one amino acid to the nitrogen atom from another amino acid. This reaction is called a dehydration synthesis reaction, and occurs in the ribosome.

A protein consists of around 250 amino acids linked together in this fashion and can be viewed as a flexible chain. Amino acids interact with each other and with the water surrounding them through their side chains in several ways, and these interactions can play a role in the stability of proteins. Below I list the main four, ordered by their importance in protein folding and aggregation.

- The main focus of this thesis, the hydrophobic effect. Some amino acids have a hydrophobic side group, and are incapable of forming hydrogen bonds. This means that they can not form favorable interactions with other amino acids or with water. In general, these amino acids are therefore pushed to the inside, or "core" of the protein (see Figure 1.1). This hydrophobic "force" is the main driving force of protein stability (Branden and Tooze, 1998). Note that strictly speaking, the hydrophobic force does not really belong in this list, since it is not an interaction between single atoms, but an emergent property of the electrostatic interactions and hydrogen bonds between water molecules, and the lack of those interactions for hydrophobic particles. However, in many simulations an implicit solvent is used, so a correction is necessary to correct for the interactions in bulk water.

- The interaction of an oxygen-molecule on an amino acid with a hydrogen on another amino acid or water molecule, a hydrogen bond. Hydrogen bonds play an important role in the stability of all proteins.

- Disulphide bonds. Disulphide bonds are a special bond, that can only be formed by cysteine residues. This is due to the sulphur atom at the end of their side chain, which can form a very strong bond with another sulphur atom. It should be noted that not all proteins contain disulphide bonds.

- Electrostatics or charged interactions. Some amino acids have charges, and are attracted to oppositely charged amino acids (Baldwin, 2007). This is another stabilizing factor in most, but not all proteins.

Free energy

The concept of free energy is used in several different ways in this thesis. In general a system will tend towards a state with the lowest free energy. Or, stated in another way, the state with the lowest free energy has the highest probability. The reason we work with a free energy in stead of directly with probabilities, is that the free energy can be more easily related to the enthalpy and entropy. These two thermodynamic properties can be estimated from simulations and determine the probability a certain state has to occur. The exact relation between the free energy, enthalpy and entropy is given below.

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where ΔG is the free energy, ΔH is the enthalpy, T is the temperature and ΔS is the entropy. If $\Delta G < 0$, the state with the lowest free energy will be occupied more often than the state with the highest free energy. The ratio of the probability of the system to be in different states, can be expressed by:

$$\frac{P_1}{P_2} = e^{\frac{-\Delta G}{k_B T}} \tag{1.2}$$

We apply this concept in several different ways in this thesis. Initially, simply to protein folding and the difference in free energy between the folded and the unfolded state. Typically, the folded state of a protein has a lower free energy than the unfolded state in physiological conditions. This is due to enthalpic contributions of hydrogen bonds, electrostatic interactions, and hydrogen bonds. The hydrophobic effect, which is partly entropic and partly enthalpic results from a combination of these contributions and the geometry of water molecules around hydrophobic particles. Due to these factors, the entropy (S) of a folded protein is generally unfavorable compared to the unfolded state, whereas the enthalpy of the folded state is favorable. One can see from equation (1.1) that if one increases the temperature, the entropy becomes more important. That means that at higher temperatures, ΔG becomes positive and a protein will unfold. Some proteins also cold denature, which is counterintuitive and conflicts with eqn. (1.1). We investigate this phenomenon further in Chapter 3.

We also use free energies to estimate the interactions involving amino acids. Amino acids interact with each other and with water. The pairwise potential used in our simulations were derived previously using an adapted form of eqn. (1.2) (Miyazawa and Jernigan, 1985b). Over a large database of native protein structures, we calculated the probability of observing different types of amino acids in close proximity. A similar approach was introduced by (Abeln and Frenkel, 2011) for the interactions of amino acids with water. Because both the enthalpic and entropic component of the hydrophobic effect change with temperature, the interaction free energy of an hydrophobic amino acid with water changes with temperature. To find the free energy for a larger temperature range, we adapt the above approach so that it applies even when the temperature is higher or lower than normal. See chapter 3 for more details.

We also use the abstraction of a free energy landscape. The concept of a free energy landscape is perhaps one of the most famous abstractions used in protein folding. It not only considers the free energy for the folded state and the unfolded state, but also the transition between the folded and the unfolded state (Figure 2.1). Since the transition between unfolded state to the folded state can be complicated, and is not a single path but a collection (also know as an 'ensemble') of paths, we need to define an order parameter. The order parameters we used are internal contacts and native contacts for protein folding, and external contacts and hydrogen bonds for protein aggregation.

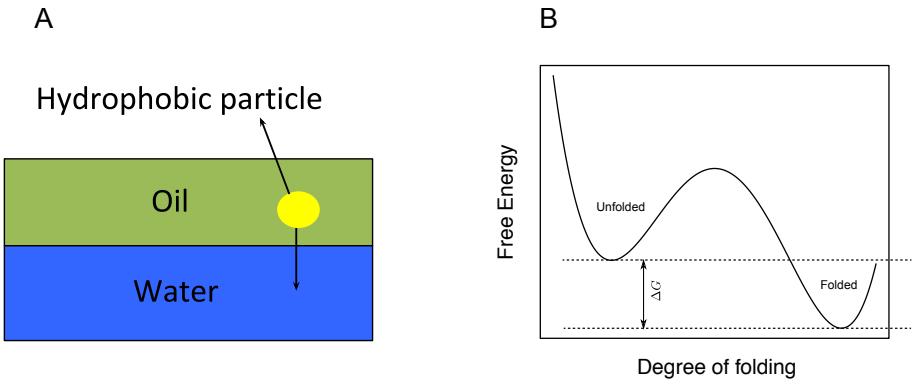

FIGURE 1.2: **Two applications of free energy used in this thesis**. In panel (A) the concept of a transfer free energy is illustrated. For the purposes of this thesis, we define the transfer free energy as the energy required to transfer a particle (in our case an amino acid) from water to oil. It can also be estimated using the partition coefficient: the number of particles in water divided by the number of particles in oil. It is estimated using a statistical approach in chapter 2 for single amino acids and we compared with theoretical calculations for purely hydrophobic particles in chapter 3. In panel (B) we show the concept of a free energy landscape. In reality, the picture is a lot more complex: There are multiple pathways from the unfolded to the folded state. Moreover, it appears that in typical conditions an amyloid fibril is actually the most stable state.

Heat capacity

The heat capacity is defined as the amount of heat required to change the temperature by a single degree:

$$C = \frac{\delta Q}{\delta T} \tag{1.3}$$

The heat capacity of a system can be measured experimentally. It allows one to gain insight in the entropy and enthalpy of a system more directly than by using computer simulations.

In a finite system, a phase transition or a reaction is characterised by a sharp peak in the heat capacity as the system is heated. This is due to the transition from an enthalpically favourable to an entropically favourable state, which requires the addition of heat before the temperature is increased. This is also seen for the unfolding of proteins, and can be used to obtain the enthalpy of folding. There is one subtle issue with these estimations: the heat capacity of the unfolded protein is higher than the heat capacity of a folded protein. Moreover, the heat capacity of a folded protein increases linearly with

temperature. In the past, this problem has been addressed by a fitting procedure that estimates the heat capacity of the folded and the unfolded proteins (Prabhu and Sharp, 2005). Others have attempted to find predictions for the heat capacity of folded protein based on several characteristics of the protein. This thesis presents a theoretical model that predicts a relation between the heat capacity and the temperature dependence of the hydrophobic effect. In chapter 4, we show that a prediction method based on this theoretical model improves upon this work. This method, HSApred, is publicly available through a web server.

A brief summary of simulation methods

A quantitative estimate of the interactions that are important in protein folding should, in theory, allow us to make predictions on the protein structure, interactions, free energy of folding and tendency to aggregate. This requires a method that takes all the interactions on the level of individual atoms and the starting coordinates of a specific protein structure, and provides as output predictions on the protein level. Currently, there is no perfect way to compute accurate predictions using only these atomistic interactions. Many methods exist, each with their own advantages and drawbacks. In this thesis the focus will be on a commonly used class of methods that have a physical basis, simulation methods.

To use the interactions described above in a simulation, they need to be described in a "potential" or a "forcefield". Generally, interactions between atoms can be divided into electrostatic reactions and van der Waals interactions. These interactions can be easily converted to forces if the distance between atoms is known. In a *molecular dynamics* simulation, the forces acting on each atom are calculated, and subsequently converted to acceleration using Newton's laws. This process is repeated over a large number of time steps. In a *Monte Carlo* simulation, there is no notion of velocity or acceleration of atoms. Instead, random moves are made and accepted if the energy is decreased. If the energy increases, the moves are accepted with a certain probability, calculated through the Boltzman criterium. Both methods have benefits and drawbacks. In general, a model that uses a Monte Carlo simulation is easier to implement, and allows for potentials that are not continuous. Molecular dynamics has the advantage that one can study the kinetics or movement of a protein. In this thesis the Monte Carlo algorithm is used.

Calculating the interactions between all atoms is very time consuming. Both molecular dynamics and Monte Carlo simulations require many steps, and for each step all interactions are required. This means that a simulation over a reasonable timescale takes a very long time. Using a normal desktop computer, simulations of around a microsecond (0.000001 seconds) in simulated time take about a month on a normal desktop CPU. Using specialised hardware, millisecond (0.001 seconds) simulations can be done for small systems, and take about a month as well. A millisecond simulation is long enough to simulate the folding

of a small (35 amino acids), quickly folding protein, FiP35. However, for larger proteins the simulation becomes slower and the folding time increases. Simulating protein aggregation with atomistic forcefields is completely unfeasible, since the phenomenon occurs over the course of hours to years.

Coarse grained simulations

To improve the computational requirements many ways of simplifying the coarse grained interactions have been proposed. Since proteins are composed of 20 different types of amino acids, a common simplification is to capture the interactions between amino acids, as opposed to the interactions between the atoms in amino acids. The number of pairwise interactions in a systems grows quadratically with the number of particles, so this means a drastic reduction in the number of interactions. Another common approximation is to replace the water molecules with an implicit solvent model. An implicit solvent model has no particles, but uses an effective interaction between water and amino acids or atoms if no other particles are nearby. Known coarse grained simulation models include:

- The MARTINI force field. Reduces 4 atoms to a single particle. Originally developed to simulate lipid and surfactant systems, currently also used for protein protein interactions (Marrink et al., 2007; Monticelli et al., 2008; Periole et al., 2012; May et al., 2014).
- The tube model. Has an interaction potential for amino acids, implicit solvent. Used for folding simulations (Kukic et al., 2015)
- Lattice models. Single interaction potential for amino acids, implicit solvent. The simulation uses a cubic lattice. (Abeln and Frenkel, 2011; van Dijk et al., 2016c)

In this thesis, a coarse grained lattice model is used. Amino acids are represented by a single pseudo-atom, which can interact with water and with other amino acids. This simplification drastically reduces the required simulation times. On my desktop system, a protein can fold in around fifteen minutes. Unfortunately, the accuracy of this simulations are reduced quite drastically. So much so, that it is no longer possible to make predictions for specific proteins using this method. However, one can still make statements about the factors in the potentials that are important for all proteins.

While the model has its limitations because it is simpler and in most cases less accurate than more fine-grained models, it is also easier to understand the results it produces. We find that incorporating the hydrophobic effect in a lattice model allows us to estimate what relative contributions of the different energies are required to reproduce known behavior. Our results are consistent with experimental measurements of the hydrophobic effect on the protein level and with predictions based on physical theory on the level of individual atomistic interactions. This allows our model to "bridge" the gap

between simulations on a global level, and known results on the atomistic level. Our model also yields the novel prediction that the slope of the heat capacity of a protein with regard to temperature is directly related to its hydrophobic surface area, allowing us to predict properties of real proteins.

The thesis is divided in five major parts:

- Determining how the hydrophobic interactions change with temperature (Chapter 2, beginning of Chapter 3)
- Simulating protein folding while taking into account these temperature dependent changes, and using these simulations along with theoretical calculations to make predictions for trends that should hold for all proteins (Chapter 3).
- Testing these predictions for a real dataset, and making the results available for the public (Chapter 4)
- Using bioinformatics methods to help these predictions when no structure is available (Chapter 5)
- Exploring the implications of the temperature dependence for protein aggregation (Chapter 6).

Chapter 2

The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures

Based on the publication:

Erik van Dijk, Arlo Hoogeveen, and Sanne Abeln (2015). "The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures". In: *PLoS Comput Biol* 11.5, e1004277. DOI: 10.1371/journal.pcbi.1004277. URL: http://dx.doi.org/10.1371\%2Fjournal.pcbi.1004277

Abstract

The hydrophobic effect is the main driving force in protein folding. One can estimate the relative strength of this hydrophobic effect for each amino acid by mining a large set of experimentally determined protein structures. However, the hydrophobic force is known to be strongly temperature dependent. This temperature dependence is thought to explain the denaturation of proteins at low temperatures. Here we investigate if it is possible to extract this temperature dependence directly from a large set of protein structures determined at different temperatures. Using NMR structures filtered for sequence identity, we were able to extract hydrophobicity propensities for all amino acids at five different temperature ranges (spanning 265-340 K). These propensities show that the hydrophobicity becomes weaker at lower temperatures, in line with current theory. Alternatively, one can conclude that the temperature dependence of the hydrophobic effect has a measurable influence on protein structures. Moreover, this work provides a method for probing the individual temperature dependence of the different amino acid types, which is difficult to obtain by direct experiment.

Introduction

When a protein folds, hydrophobic amino acids get buried inside the protein to form a hydrophobic core. Inside this core the hydrophobic side chains are shielded from the water. The tendency of hydrophobic groups to cluster together when they are put into water - or the hydrophobic effect - is the most important driving force in protein folding. Note that there are several factors that contribute to the overall stability of a folded protein: for example the formation of hydrogen bonds between backbone atoms (secondary structure) and side chains; the formation of salt bridges between charged amino acids and the burial of hydrophobic side chains upon folding. It is thought that this hydrophobic force gives the single largest contribution to the stability of most protein folds (Baldwin, 2007). Moreover, the positioning of hydrophobic clusters in the sequence may affect the folding pathway and dynamics e.g. (Wood et al., 2011; Zarrine-Afsar et al., 2008). Note that these stabilising forces are partially compensated by the decrease in chain entropy upon folding.

Hydrophobicity is a result of the collective behaviour of the water molecules and 'oily' groups. In essence the water-hydrophobe interface is unfavourable compared to water-water or hydrophobic-hydrophobic interactions. The free energy difference upon burial of hydrophobic groups is partially entropic and partially enthalpic, causing a distinct temperature dependence (Widom, Bhimalapuram, and Koga, 2003a; Chandler, 2005). Even though the exact molecular cause for these enthalpic and entropic contributions is the focus of active research (Rezus and Bakker, 2007; Hazy et al., 2011) and can change depending on the type of protein (Hazy et al., 2011), the resultant temperature dependence can be measured experimentally for several different non-polar substances (Rettich et al., 1981; Kim and Szoka, 1992). From such measurements, models and theory we know that the hydrophobic force peaks between 30-80 °C and becomes weaker at both lower and higher temperatures, see Fig. 2.1A.

Since hydrophobicity is such a large contributor to protein stability, the temperature dependence of the hydrophobic effect has important consequences. Firstly, some proteins do not only unfold at high temperatures, as can be explained through the entropy of the chain, but also at low temperatures (cold denaturation) (Vajpai et al., 2013). This effect is thought to be a consequence of hydrophobicity becoming weaker at low temperatures (Dias et al., 2008). Secondly, alternate states of intrinsically disordered proteins may become more favourable at different temperatures due to this effect (Uversky, Li, and Fink, 2001). Thirdly, protein-protein and protein-substrate interactions – if dominated by hydrophobic interactions – may also be sensitive to temperature changes.

It is essential to quantify the temperature dependence if one wants to model and predict the stability of folded proteins and protein interactions over a large range of temperatures. For industrial purposes, proteins or enzymes that can be active over a wide temperature range are of crucial importance. To achieve this, proteins from species that live at extreme temperatures, thermophiles and

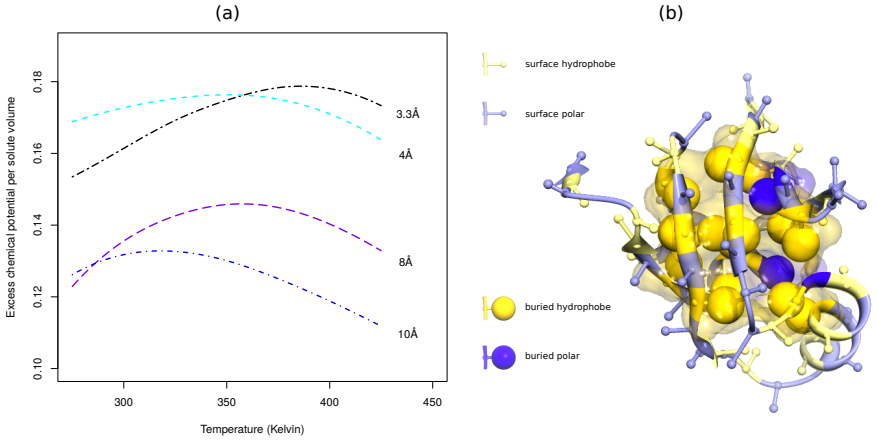

FIGURE 2.1: **Length scale dependence of hydrophobic effect from calculations by Huang and Chandler (Huang and Chandler, 2000) (A)**. The cost of making a cavity in the water with a radius of the given size against temperature is plotted. The position of the maximum depends on the size (radius) of the solute. Small solutes with a radius of 3.3 Å have a peak at around 70 °C, whereas larger particles with a radius of 10 Å have a peak around 40 °C. **An example protein structure: PDB-ID: 2K5I (B).** We estimate free energies of transfer from the hydrophobic core to the surface of the protein by comparing the number of hydrophobic amino acids on the surface (small yellow spheres), to the number of buried hydrophobics (large yellow spheres), to the number of polar amino acids on the surface (small blue spheres) and to the number of buried polar amino acids (large blue spheres).

psychrophiles, have been used and adapted extensively for biocatalysis (Frock and Kelly, 2012; Tajima et al., 2013). Understanding and quantifying the hydrophobic temperature dependence for specific amino acids is essential if one wants to predict thermostability of proteins.

Earlier, (Folch, Dehouck, and Rooman, 2010) showed that temperature dependent pairwise potentials for amino acids can help to predict the melting temperature of homologous pairs of proteins. More recently, this study was extended to also predict stability at low temperatures (Pucci and Rooman, 2014). In this work we focus on the temperature dependence of the effective interactions between hydrophobic amino acids and water.

Even though this temperature dependence has important consequences, it is often not considered due to practical concerns. The temperature dependence is typically not included in interaction potentials for protein structure prediction or coarse grained simulations; such potentials do not model the water molecules

explicitly or in enough detail to capture this effect. It is difficult to measure the temperature dependence for specific amino acids by experiments, under physically relevant conditions. In other words, it is difficult to measure the difference in free energy between the folded and unfolded chain for separate amino acids. In this work we show that it is possible to obtain this temperature dependence for specific amino acids by mining a large set of protein structures resolved by Nuclear Magnetic Resonance (NMR).

Physically or chemically relevant quantities can be obtained by averaging over a large set of structures. For example, specific bond lengths, the most favourable dihedral angles or approximate hydrophobicities for different amino acid types can be obtained by taking an ensemble average over a set of protein structures. More specifically, hydrophobicity scales for the different amino acid types may be obtained using physicochemical properties (Kyte and Doolittle, 1982), or by calculating how often we find each residue type exposed to the solvent at the surface of a protein (Chothia, 1976a; Kyte and Doolittle, 1982; Rose et al., 1985; Shaytan, Shaitan, and Khokhlov, 2009). Different approaches give slightly different results – and a somewhat different ranking between the residues – but do agree overall. Hydrophobicity scales are useful for a wide range of problems involving structure prediction: from predicting the severity of a mutation to disorder prediction and full structure prediction e.g. (Venselaar et al., 2010; Floris et al., 2011; Oldfield et al., 2005; Cilia et al., 2013; Zhou and Zhou, 2002; Buchete, Straub, and Thirumalai, 2004).

Estimates for pairwise free energies between amino acid types have been obtained by mining protein structures. A pairwise interaction potential may be calculated by counting the number of contacts made between different types of amino acids (Miyazawa and Jernigan, 1985a; Betancourt and Thirumalai, 1999; Folch, Dehouck, and Rooman, 2010). More recently, this method has been further developed to allow the extraction of interactions between the solvent and the different types of residues, as well as the pairwise interactions (Abeln and Frenkel, 2011). Knowledge-based amino acid pair-potentials are used in structure prediction (Shen and Sali, 2006), coarse-grained protein simulations (Coluzza, Muller, and Frenkel, 2003; Coluzza and Frenkel, 2007; Ni et al., 2013a; Abeln et al., 2014a) and protein-protein docking methods (Halperin et al., 2002). Recently, a knowledge based amino acid pair potential with a temperature dependence has also been used to predict the thermostability of proteins (Pucci and Rooman, 2014).

In this work, we estimate the hydrophobic effect as the free energy cost for transferring a hydrophobic amino acid from the core of the protein to the water exposed surface, see Fig. 2.1B. We use three distinct approaches to estimate these transfer free energies. Firstly, we use a previously validated approach to derive a statistical pair potential between amino acids to extract free energy estimates for the hydrophobic interaction. This *contact* based method has been shown to yield hydrophobicity estimates that give physically realistic results upon simulation. Secondly, we use a more direct approach that calculates propensities for *surface* accessibility for each of the amino acids; this method is

similar to other approaches that derive knowledge based hydrophobicity scales (Chothia, 1976a; Kyte and Doolittle, 1982; Rose et al., 1985). Thirdly, we use an *area* based approach that considers the amount of exposed surface area per amino acid. The three approaches give similar results, and show significant temperature dependence for hydrophobic amino acids in line with expectations from theory and measurements on small hydrophobic particles.

Results/ Discussion

In order to extract the hydrophobic temperature dependence from experimentally determined protein structures, it is important to choose the set of structures carefully. Firstly, we explored the contents of the Protein DataBank (PDB), (Berman et al., 2000a), containing over 96k structures. Fig. 2.2 shows the temperature distribution of available protein structures determined by X-ray crystallography and nuclear magnetic resonance (NMR). For this study we only use structures determined by NMR, as these experiments can be performed on soluble proteins at the temperature range of interest for the hydrophobic effect. This makes it possible to probe temperature dependent effects in proteins; for example a temperature induced transition (Thiriot, Nevzorov, and Opella, 2005) and cold denaturation (Vajpai et al., 2013) have been observed using this technique.

In order to obtain estimates for the solvation free energies of different types of amino acids at different temperatures, we divided the data into five temperature bins, see Fig. 2.2 and Table 2.1. The bins were chosen symmetrically around the peak at room temperature (300 K), to balance the number of structures in each bin.

TABLE 2.1: **Selected protein structures**

temperature range	chains in PDB	chains select-25	chains after filters
265-290	1421	259	207
291-296	1440	378	344
297-299	4689	1095	1033
300-305	1864	618	560
306-340	1361	470	412

The number of NMR chains as present in the PDB before and after filtering is shown for different temperature bins (in kelvin). At the first stage of filtering, sequence bias was removed using PDB-select-25. At the last filtering step, chains were removed when they were not compatible with DSSP or when they had multiple and different acquisition temperatures.

We set out to explore if we can observe the temperature dependence of the hydrophobic effect by analysing this filtered set of protein structures. Protein structures determined by NMR at different temperatures were used to obtain

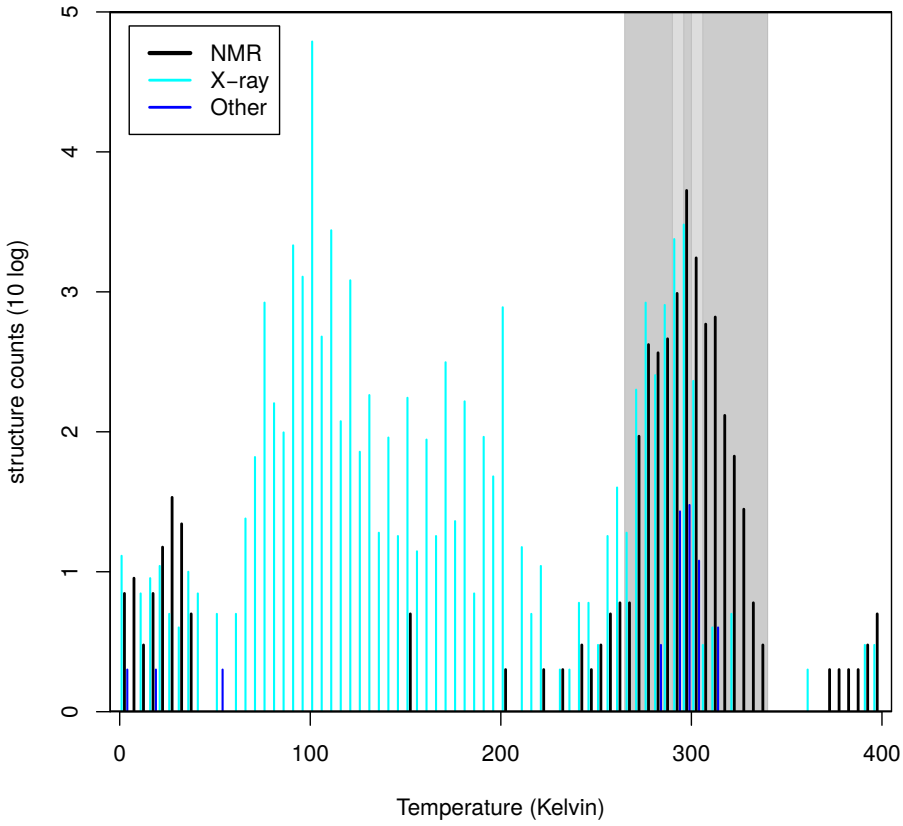

FIGURE 2.2: **Distribution of temperatures at which experimental protein structures were resolved.** All acquisition temperatures of structures as of April 2014 available in the PDB are shown. The 80,662 X-ray diffraction structures are centred around 100 K, while the 10,969 NMR structures show a peak at room temperature (300 K). Note that the small peak of NMR data just above absolute zero may be temperatures entered in celsius instead of kelvin; this data is not used in this study. Temperature bins, as given in Table 2.1, are indicated in different shades of grey.

free energy estimates for the transfer of amino acids from the core of the protein to the surface. Under the assumption of random mixing, the transfer free energy estimates can be estimated through statistical methods (Miyazawa and Jernigan, 1985a; Betancourt and Thirumalai, 1999; Folch, Dehouck, and Rooman, 2010; Abeln and Frenkel, 2011). We investigate three methods, 1) a

TABLE 2.2: **Amino acid class definition**

Class	Amino Acids
Hydrophobics	ALA, ILE, LEU, MET, VAL
Aromatics	HIS, PHE, TRP, TYR
Charged	ARG, ASP, GLU, LYS
Polar	ASN, GLN, SER, THR
Other	CYS, GLY, PRO

contact based calculation which has been shown to give a reasonable attraction (Abeln and Frenkel, 2011), 2) a direct calculation of propensities to *surface* exposure 3) an *area* based calculation that incorporates the accessible surface area in a continuous measure of hydrophobicity, see Methods for details.

Firstly, we investigate whether the raw free energy estimates are dependent on the temperature. To further increase the statistical accuracy, amino acids are divided into five classes: hydrophobic, charged, polar, aromatic and other, see Table 2.2. Fig. 2.3 shows a surprisingly clear temperature dependence for the different hydrophobic amino acids: at lower temperatures the hydrophobic effect becomes weaker. This is in line with expectations from experiments and theory (Widom, Bhimalapuram, and Koga, 2003a; Chandler, 2005). The results for the area based potential are very similar to the results of the contact based potential (see original publication, supplemental Figures S7-S14).

To test if this temperature dependence is indeed significant, we resampled the protein structures using random temperature labels. From this procedure p-values were calculated to determine the significance of the free energy difference. Table 2.3 shows the difference in transfer energy ($\Delta\Delta G$) and p-values between the lowest temperature bin (265-290K) and room temperature (297-299K). Clearly, the temperature trend for the hydrophobic residues is significantly stronger than one would expect from random fluctuations. The standard error to the mean is estimated from the deviations in the potential obtained as indicated in the results by splitting the data set into five parts and recalculating the potentials for each part.

Fig. 2.3 also shows that the surface based potentials give larger absolute differences in free energies than the contact based potentials. This can most likely be explained by the strict cutoff (7% accessible surface area) in the surface based potential compared to the more gradual calculation of the contact based potential; charged and polar amino acids are rarely entirely buried and give therefore a very strong signal for the surface based measure. The relative hydrophobicity, however, is consistent between the three methods, showing our results are qualitatively independent of the method of derivation for the potential.

The results in Fig. 2.3 show a slight temperature dependence for charged (and polar) amino acids. For the surface based potential, however, this effect is not significant (Table 2.3).

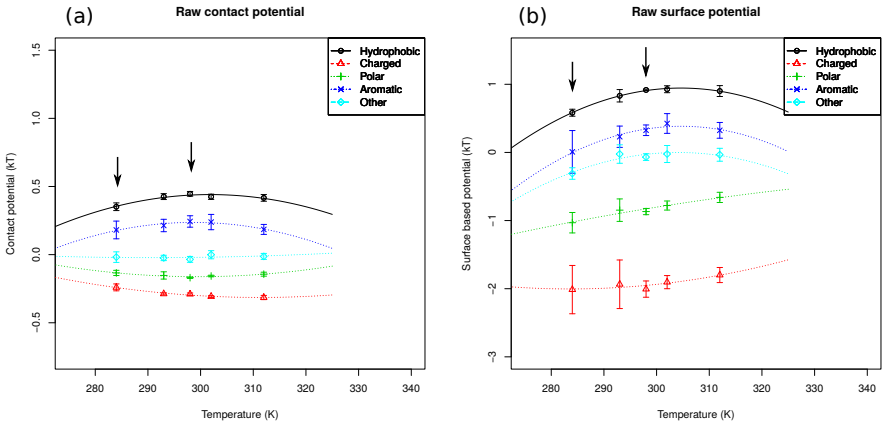

FIGURE 2.3: **Raw free energies of transfer for classes of amino acids.** Contact based (A) and surface based (B) free energies are shown for different classes of amino acids. Points show the free energy estimates for each temperature bin, lines are fitted with a parabola, consistent with the potentials found in (Huang and Chandler, 2000). Arrows indicate the bins used to test the significance of the temperature dependence.

TABLE 2.3: **Significance of hydrophobic temperature dependence pooled**

amino acid class	p-value contacts	p-value surface	$\Delta\Delta G$ contacts	$\Delta\Delta G$ surface
hydrophobic	< 0.01	< 0.01	0.10	0.32
polar	< 0.01	0.23	-0.05	0.13
charged	< 0.01	0.80	-0.06	-0.04
aromatic	0.04	< 0.01	0.06	0.32
other	0.32	< 0.01	0.02	0.41

The difference in free energy estimates ($\Delta\Delta G$) between the lowest temperature bin (265-290K) and room temperature (297-299K) is shown together with its significance (p-value) for each class of amino acids. The significance was tested using a resampling procedure. The amino acids are pooled according to defined classes; the free energy estimates are not reference corrected.

Our transfer free energy estimates are calculated under the assumption of a random mixing model; this provides us with *relative* transfer free energies for each type of amino acids. This means it is not trivial to compare the free energy differences between different temperature bins. The temperature

dependence of the hydrophobic residues could cause the shift of the polar and charged amino acids. In order to enable comparison at different temperatures, we set a reference state for the free energy estimates. The reference state is an important part of the potential, and can determine the accuracy of a potential in structure validation (Shirota, Ishida, and Kinoshita, 2009).

As we are here particularly interested to compare the transfer free energies between different temperatures it is desirable that our reference does not have any temperature dependent interaction with the solvent. Betancourt and Thirumalai (1999) and Buchete, Straub, and Thirumalai (2004) use Threonine, a small water-like polar amino acid, as a reference in the calculation for their amino acid pair-potential. In our case, as the number of structures available is limited, choosing a single amino acid as reference will propagate noise through the results. Instead, we pool all the charged and hydrophilic amino acids for each temperature bin, and use those as a reference potential (see Table 2.2). Even though it is known that polar and charged residues can have a temperature dependent interaction with the solvent and that this interaction can have consequences for protein structure and stability (see for example Refs. (Wuttke et al., 2014; Chamberlin, Cramer, and Truhlar, 2006)), comparing *raw* estimates (Fig. 2.3) with *reference corrected* estimates (Figures S1 and S4, original publication) shows that this correction does not change the relative trends, see Methods for further details.

Fig. 2.4 shows estimates for the corrected transfer free energies for all hydrophobic and aromatic amino acids individually, with the polar and charged amino acids as a reference. Results for all amino acids, with and without reference correction are shown in the supplemental information of the original publication, Figures S2, S3, S5 and S6. The hydrophobicity becomes weaker at lower temperatures, showing the results from the 'raw' estimates hold up. Again, the significance of the temperature dependence of each hydrophobic amino acid type is examined. For almost all hydrophobic amino acids the free energy estimates have a significant temperature dependence (Table 2.4). Note that the correction to a reference of polar and charged amino acids was also performed in the resampling procedure to obtain statistical significance

Fig. 2.4 also shows that the estimated transfer free energies show a very similar trend with respect to temperature to those that have been measured for hydrophobic particles (Widom, Bhimalapuram, and Koga, 2003a) or obtained by calculation according to LCW-theory (Lum, Chandler, and Weeks, 1999; Huang and Chandler, 2000). For clarity, we fitted parabolas through the estimated transfer free energies, which is a reasonable approximation for trends calculated from theory and observed in experiment (see Original publication, Figure S15). It can be observed that the free energies for the hydrophobic amino acids show a maximum of around 310-350 kelvin for both the surface and contact based free energy estimates; this is slightly lower than what is expected from theory (see for comparison Fig. 2.1A)

Due to the lack of data at higher temperatures ($T > 320K$), it is difficult to estimate a precise maximum for the transfer free energies. Nevertheless,

TABLE 2.4: **Significance of hydrophobic temperature dependence**

amino acid (class)	p-value contacts	p-value surface	$\Delta\Delta G$ contacts	$\Delta\Delta G$ surface
ALA	0.03	< 0.01	0.12	0.38
CYS	< 0.01	< 0.01	0.32	0.67
GLY	0.42	0.12	0.03	0.11
ILE	< 0.01	< 0.01	0.20	0.19
LEU	< 0.01	< 0.01	0.19	0.32
MET	0.68	0.09	0.03	0.19
PHE	< 0.01	< 0.01	0.23	0.36
PRO	0.62	< 0.01	0.07	0.46
TRP	0.23	0.21	0.15	0.16
TYR	0.34	0.08	0.10	0.19
VAL	0.02	< 0.01	0.15	0.16

The difference in free energy estimates ($\Delta\Delta G$) between the lowest temperature bin (265-290K) and room temperature (297-299K) is shown together with its significance (p-value) for each class of amino acids. The significance was tested using a resampling procedure. The hydrophobic and aromatic amino acids are shown and are reference corrected with respect to the charged and polar amino acids.

an interesting trend may be observed from Fig. 2.4. Larger amino acids, for example Tryptophan, have a maximum at lower temperatures compared to smaller amino acids such as Alanine. Again, this trend is consistent with theory and experiments (Huang and Chandler, 2000), where the transfer free energy of larger particles shows a maximum at lower temperatures.

Overall, we can conclude that the temperature dependence of the hydrophobic effect has a measurable influence on protein structures determined by NMR. The effect we find appears to be on the right order of magnitude in comparison with theory for the hydrophobic effect and known cold denaturating behaviour of proteins (see S2 text). The results show that structures determined at lower temperature have more exposed hydrophobic surface area. This suggests that at these temperatures the structures already become more open, as has been observed for some specific proteins (e.g. (Gast et al., 1994)). It would be very interesting to investigate if these low temperature structures are more flexible and dynamic than the same structures obtained at room temperature.

Conclusion

In this work we set out to investigate whether the hydrophobic temperature dependence could be obtained by mining a large set of protein structures resolved by NMR. We used a contact based, an area based and a surface based approach to obtain free energy estimates for the transfer of an amino acid out

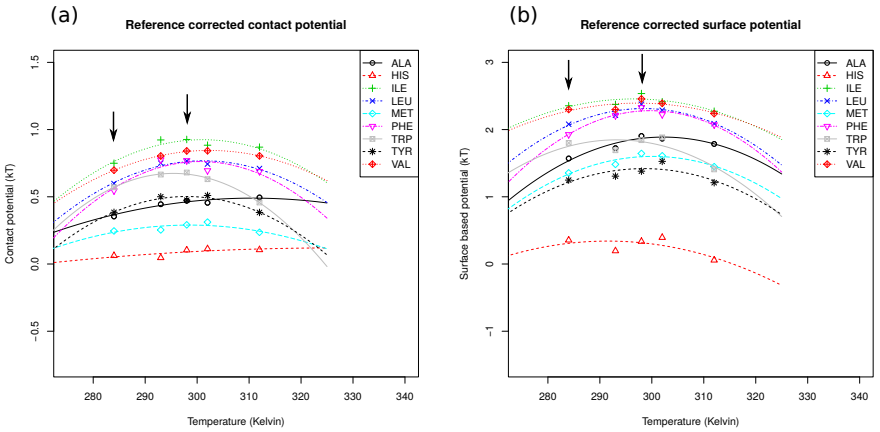

FIGURE 2.4: **Reference corrected free energies of transfer for hydrophobic amino acids.** Contact based (A) and surface based (B) free energies are shown for hydrophobic and aromatic amino acids. The free energies are corrected by setting a reference of the polar and charged amino acids. Points show the free energy estimates for each temperature bin and lines are fitted with a parabola. Arrows indicate the bins used to test the significance of the temperature dependence.

of the hydrophobic protein core onto the water exposed surface. We find a surprisingly clear trend for the free energy estimates with respect to the temperature: the hydrophobic effect becomes weaker at lower temperatures, as is expected based on theory, simulations and experiments. Alternatively, one can conclude that the temperature dependence of the hydrophobic effect has indeed a measurable influence on protein structures. Despite the sparseness of the data, and the inconsistencies in reporting of experimental temperatures, we find that the observed trend holds and is significant regardless of the precise method used to estimate the transfer free energies, the specific groupings of amino acids or the chosen reference.

Methods

Data collection

The temperature (in kelvin) at which the experiment is performed can be found in the mandatory 'acquisition data' section of PDB files. Several filters were applied. Some structures were filtered out because no temperature was entered or because they were given several temperatures from multiple data collection sessions. In order to get representative statistics for amino acid composition,

it is important to remove any bias in the PDB for large sequence families. To take out this redundancy we used PDB filter-select 25% (Hobohm et al., 1992; Hobohm and Sander, 1994; Griep and Hobohm, 2010). Table 2.1 shows the number of remaining structures in each bin after these filtering steps in each temperature bin. A few further PDB files had to be removed due to their incompatibility with DSSP. After these steps, each PDB-file was split into multiple models, and the accessible surface area was determined using DSSP for each model. For each residue in the protein chain, the average accessible surface area over all models was used. The final counts for each PDB-structure are shown in S1 data, original publication. The format is explained in S1 Text.

Calculation of contact based potential

To obtain estimates for the free energies of transferring specific amino acid types from the outside of the protein to the hydrophobic core, we used two approaches. The first approach is based on *contacts* between amino acids, and between amino acids and the solvent as in the work of Abeln and Frenkel (Abeln and Frenkel, 2011). This potential has been shown to give an appropriate distinction between the protein core and surface by simulation. The second approach uses the presence or absence of amino acids on the *surface* of the protein, providing a more direct way to obtain the hydrophobicty of each amino acids.

In the contact based approach, we calculate knowledge-based pair-potentials over the set of structures described above. The free energy estimates $\epsilon_{i,j}$ between amino acid types i and j can be calculated as:

$$\epsilon_{i,j} = -kT \ln \left(\frac{c_{i,j}}{\omega_{i,j}} \right) \tag{2.1}$$

where $c_{i,j}$ are the number of contacts between amino acids type i and j, and where $\omega_{i,j}$ is the expected number of contacts. Note that here we are specifically interested in the case where one of the interaction partners is the solvent, i.e. $\epsilon_{i,\text{solvent}}$.

We can calculate the expected number of contacts, $\omega_{i,j}$, by considering the distribution of the amino acid types i and j in the set of protein structures:

$$\omega_{i,j} = \frac{n_i q_i n_j q_j}{\sum_k q_k n_k} \tag{2.2}$$

here $n_i q_i$ is the total amount of contacts for type i, where n_i is the number of amino acid of type i and q_i is the coordination number, which we set to 4 for all amino acids to remain consistent with Abeln and Frenkel (Abeln and Frenkel, 2011). Note that the sum in denominator loops over all the amino acids and water (k). In practise the total number of contacts for an amino acid type $n_i q_i$ can be calculated directly from the data.

The number of water contacts is estimated through the size of the surface accessible area for a residue as calculated by DSSP (Kabsch and Sander, 1983). Note that for the water contact points, we do not consider real water molecules, but a surface area similar to the size of an amino acid. We estimate the number of contacts as the product between $q = 4$ and the fraction of exposed surface area α_r for residue r. Hence, based on the assumption that a residue can interact with four other residues, water contact points can be created. The fraction of exposed surface area, α_r, is given by:

$$\alpha_r = \frac{S_r}{\max\left\{S_{a(r)}\right\}} \tag{2.3}$$

S_r is the solvent accessible area, calculated with the DSSP program, and $a(r)$ is the amino acid type of residue r; $\max\left\{S_{a(r)}\right\}$ is the maximum accessible area in an unfolded chain for that amino acid type.

Calculation of surface based potential

An alternative measure for hydrophobicity can be obtained by calculating the propensity for an amino acid to be on the surface. Classic amino acid propensities, which are for example used to describe the affinity for a certain secondary structure type, can be calculated through a simple ratio of fractions e.g. Chapter 12 of Ref. (Zvelebil and Baum, 2008). Here we use the structural classes buried and non-buried. To decide whether a residue (r) is buried, we use a cut-off: $\alpha_r < 7\%$ (Hubbard and Blundell, 1987). We can calculate the propensity (P) for amino acids to be buried as:

$$P_{a,b} = p_{a,b}/p_b \tag{2.4}$$

where $P_{a,b}$ stands for the propensity for an amino acid type, a, to be buried as indicated by the subscript b. Translating this into counts yields:

$$p_{a,b} = \frac{N_{a,b}}{N_{a,b} + N_{a,nb}} \tag{2.5}$$

where $N_{a,b}$ is the total number of amino acids of type a that are buried, and $N_{a,nb}$ is the total number of amino acids of type a that are non-buried. Similarly,

$$p_b = \frac{N_b}{N_b + N_{nb}} \tag{2.6}$$

where N_b is the total number of buried amino acids, and N_{nb} is the total number of amino acids that are not buried.

When propensities are used to estimate transfer free energies, through $\Delta F_{a,b} = -kT\log(P_{a,b})$ it has the disadvantage that:

$$\Delta F_{a,b} \neq -\Delta F_{a,nb} \tag{2.7}$$

This can be seen by substituting the formula for $P_{a,nb}$ in the formula for the free energy, ΔF.

Here we define our propensities in an alternative way to overcome this problem analogously to the definition in Shaytan, Shaitan, and Khokhlov (2009). If we define our alternative propensities, P^*, analogous to a partition coefficient, we obtain:

$$P^*_{a,b} = \frac{p^*_{a,b}}{p^*_{a,nb}} = \frac{N_{a,b}/N_{a,nb}}{N_b/N_{nb}} \tag{2.8}$$

which does have the desired property summarized in eqn. 2.7.

Calculation of area based potential

While the contact based potential is established, some of the assumptions are particularly useful in the context of a coarse grained lattice simulation. On the other hand, the surface based potential uses the assumption that a residue is buried when less then 7% of its surface is exposed. To test the robustness of our results with regards to these assumptions, we investigated two additional potentials, based on the exposed *area*. The first one corresponds to the contact based potential, with very large (infinite) coordination numbers. This area based potential is calculating by comparing the amount of exposed surface area, S_r for an amino acid type a to that of the average amino acid.

$$C_a = -\log\left(\frac{N}{N_a}\frac{\sum_{i\in a}\frac{S_{r_i,a}}{\max(S_{a(r_i)})}}{\sum_j \frac{S_{r_j}}{\max(S_{r_j})}}\right) \tag{2.9}$$

S_r is the solvent accessible area, calculated with the DSSP program, and $a(r)$ is the amino acid type of residue r; $\max\left\{S_{a(r)}\right\}$ is the maximum accessible area in an unfolded chain for that amino acid type.

A similar potential, but scaled with the maximum solvent accessible area, is also calculated. We will refer to this potential as the scaled area based potential, $C_{a,s} = C_a \max(S_a)$. The interactions of each residue are multiplied by its maximum accessible surface area. The results for this potential are very similar. Large residues have a higher interaction score when compared to smaller residues. The results for this potential are shown in the supplemental information in the original publication, Figures S11-S14.

Significance of temperature dependence

The estimated error to the mean for each data point was obtained by splitting the data into five parts each containing an equal number of structures. The potential was recalculated for each of the five parts, and a standard deviation was calculated from each of them. This allows us to estimate a 95% confidence interval by taking two standard errors on each side of the mean. These are the error bars shown in the plots.

The significance of the temperature dependence of the potentials was determined through a resampling procedure for two different temperature bins: the lowest temperature range and room temperature. We resampled our data by shuffling the temperature labels of the protein structures and recalculating the contact based and surface based potentials for a set of 1000 random samples. P-values for the difference in hydrophobicity between the two temperature bins were determined as the fraction of resampled free energy differences that were larger in size than the original calculation.

Fitting procedure

To obtain an estimate for the temperature dependence of the potential, we need to assign a single temperature for the structures within a temperature bin. The average temperature of the structures is taken to be the temperature of the bin. A weighted least squares fitting procedure was used to fit a parabola to the potential as a function of temperature, which is a reasonable approximation to the relation found in both theory and experiment. In a weighted least squares fit, the sum $S = \sum_{i=1}^{n} w_i r_i^2$ is minimized. Here, the i indicates the index of the temperature bin, w_i is the weight, and r_i is the difference between observations and the model. The number of residues of type s in bin i was used as weight.

2.1 Supplemental information -The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures

S1 Text: Description statistics file

Raw counts for each PDB structure per amino acid can be found at `http://www.few.vu.nl/~abeln/hydrophobicT/`. Each row corresponds to a PDB structure. Each column corresponds to a feature of the PDB structure. The identifiers in de header are described below. The statistics are provided per amino acid type, which is indicated by its three letter-code. For example, Alanine is abbreviated as ALA.

- PDB_chainID: 4-Letter pdb code followed by underscore and one letter indicating which chain is used
- Temperature: Temperature in Kelvin
- XXX_bur: Number of buried residues of type XXX. Buried is defined as less than 7% solvent accessible area.
- XXX_unbur: Number of exposed residues of type XXX. Exposed is defined as more than 7% solvent accessible area.
- XXX_Rsas: The relative solvent accessible area of type XXX. Is defined as accessible_ area/max_solvent accessible area
- XXX_Choh: Number of contacts with water for a residues of type XXX. Is defined as round(4accessible area/max_solvent_accessible_area)
- XXX_Caa: Number of contacts with other residues for a residues of type XXX. Is defined as 4 - XXX_Choh.

S2 Text: Order of magnitude estimation for temperature dependence of protein stability

To determine whether our potential yields results that are of the right order of magnitude, we will consider energy contributions for a typically sized protein. A typical protein domain has around 200 residues containing roughly 60 hydrophobic residues that are buried upon folding. Stabilities range typically between 5-20 kT at room temperature. Our potentials show that, for a hydrophobic residue the effective interaction decreases by about 0.1-0.4 kT between room temperature and 5 °C. This will yield a stability between 14 kT and -19 kT for such a protein at 5 °C. We therefore estimate that some proteins may start cold denaturating just above the freezing point of water. Hence the decrease in hydrophobic effect of 0.1-0.4 kT for this temperature range appears to be consistent with cold denaturation behavior of real proteins: most

proteins sporadically cold denature above the freezing point of water (Vajpai et al., 2013).

Many factors not considered in this calculation will also influence the folding. For example, the internal contacts have a significant temperature dependence (Pucci and Rooman, 2014). Moreover, the chain entropy is less important at low temperatures, compensating partially for the lost stability of the native state. These and other factors are different for each protein and difficult to estimate.

Chapter 3

Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation

Based on the publication:

Erik van Dijk, Patrick Varilly, Tuomas P J Knowles, Daan Frenkel, and Sanne Abeln (2016c). "Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation". In: *Phys. Rev. Lett.* 116.7, p. 78101. DOI: 10.1103/PhysRevLett.116.078101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.116.078101

Abstract

The hydrophobic effect stabilizes the native structure of proteins by minimizing the unfavourable interactions between hydrophobic residues and water through the formation of a hydrophobic core. Here we include the entropic and enthalpic contributions of the hydrophobic effect explicitly in an implicit solvent model. This allows us to capture two important effects: a length-scale dependence and a temperature dependence for the solvation of a hydrophobic particle. This consistent treatment of the hydrophobic effect explains cold denaturation and heat capacity measurements of solvated proteins.

The stability of the native state of most proteins is typically dominated by interactions between amino acids and through the hydrophobic effect. The direct amino-acid interactions can be attributed to van der Waals and electrostatic forces that are mainly enthalpic in nature. By contrast, hydrophobicity is an interaction emerging from the collective behaviour of the solvent and the side chains, and is entropy dominated (Lum, Chandler, and Weeks, 1999; Huang and Chandler, 2000; Kim and Szoka Jr, 1992; Widom, Bhimalapuram, and Koga, 2003b; Chandler, 2005) at room temperature for small solutes. The enthalpic amino acid interactions remain relatively constant over the temperature range of interest, while the magnitude of the hydrophobic effect changes with temperature (Widom, Bhimalapuram, and Koga, 2003b; Huang and Chandler, 2000).

In principle, all-atom simulations could be used to disentangle the role of entropy and enthalpy in protein folding. However, fully atomistic simulations are neither simple, nor cheap - in fact, at present, such simulations are only feasible to study the folding of relatively small proteins. Moreover, a numerical study of the stability of various protein structures would require simulations over a range of temperatures. Earlier studies (Patel et al., 2007; Patel, Debenedetti, and Frank H. Stillinger, 2007; Patel et al., 2008; Dias et al., 2008; Romero-Vargas Castrillon et al., 2012; Bianco and Franzese, 2015) have shown that a temperature-dependent hydrophobic collapse (rather than folding) can be observed in a strongly coarse-grained model for small, two-dimensional protein chains. Three-dimensional models have shown similar results for homopolymers (Bianco and Franzese, 2015) and peptides (Mitsutake et al., 2004). However, these models do not fully capture folding specificity for proteins. In a recent model that does incorporate folding specificity (Davtyan et al., 2012), two specific proteins were investigated and a linear correction was added to incorporate the temperature dependence of the hydrophobic effect (Sirovetz, Schafer, and Wolynes, 2015).

In this work we present an extension of the classic protein lattice model first introduced in Ref. (Sali, Shakhnovich, and Karplus, 1994). The classic model correctly reproduces the ability of proteins to fold into a unique native structure and it exhibits denaturation upon heating due to chain entropy alone. Interactions between amino acids are estimated through the frequency of occurrence of close contacts in experimental protein crystal structures (Miyazawa and Jernigan, 1985b). In the Miyazawa and Jernigan (MJ) potential, the interactions are strictly speaking free energies that have both entropic and enthalpic components. However, in most coarse-grained simulations these effective potentials are treated as temperature-independent enthalpies (Shakhnovich, 1994; Coluzza, Muller, and Frenkel, 2003; Abeln and Frenkel, 2008; Coluzza, 2011; Abeln and Frenkel, 2011; Abeln et al., 2014b; Kukic et al., 2015). Therefore, they do not model the temperature dependence of the interactions correctly.

In order to model the temperature dependence of the hydrophobic effect, we use an extension of the MJ potential that includes specific solvent-amino acid interaction terms (Abeln and Frenkel, 2011). The derived potential is

based on a representative subset of the Protein Database (Griep and Hobohm, 2010). The hydrophobic effect is volume dominated at small length scales and surface dominated at large length scales. Our model consistently treats this length-scale dependence by dynamically classifying residues into three categories: buried, protein surface, and fully solvated. This categorization allows us to capture the length-scale dependence of the hydrophobic effect according to Lum-Chandler-Weeks (LCW) theory (Lum, Chandler, and Weeks, 1999; Weeks, Chandler, and Andersen, 1971). Our model aims to reproduce the variation in temperature dependence for different length scales of hydrophobic solutes using these implicit solvent terms.

For each residue category, the hydrophobe-water interaction is estimated by a second-order Taylor approximation to the free energy of transfer of hydrophobic particles from an oily environment to water (see Figure 3.1). We use this model to investigate three effects that are often associated with the temperature dependence of the hydrophobic effect: Firstly, denaturation upon cooling, or 'cold denaturation'. Cold denaturation conflicts with the classical view of an entropically favourable state and an enthalpically favourable native state. Secondly, the structural characteristics of the cold denatured state. Thirdly, the temperature dependence of the heat capacity. Using Differential Scanning Calorimetry (DSC) (Hallerbach and Hinz, 1999) the heat capacity of the system can be calculated as $C_P = \left(\frac{dQ}{dT}\right)_{P,N}$. The heat capacity of the system is commonly used as a well defined experimental observable to characterize the thermodynamics of the folding transition.

We simulate a protein consisting of 80 residues with Monte Carlo sampling using a classic lattice model to investigate the effect of the entropic contribution of the hydrophobic potential. To model the effective potential for hydrophobe–water interactions, we introduce the following temperature-dependent term for the surface residues (s) and the fully hydrated residues (h):

$$F_{\mathrm{hydr}} = -\alpha_s N_{\mathrm{s}}(T - T_{0,s})^2 - \alpha_h N_{\mathrm{h}}(T - T_{0,h})^2 \tag{3.1}$$

describing second order approximations to the theory of the hydrophobic effect (Huang and Chandler, 2000; Chandler, 2005) for both groups. Here, N_{s} is the number of hydrophobic residues on the surface, N_{h} is the number of hydrophobic residues that are fully hydrated and T is the temperature in reduced units. The temperature dependence of the fully hydrated (α_h, $T_{0,h}$) and surface (α_s, $T_{0,s}$) residues are set using Ref. (Chandler, 2005) (see Figure 3.1(a)). In our lattice model, we define a residue that is fully hydrated as having at least four sides exposed and for a residue that is partially solvated as having at least one, and no more than three sides exposed to the solvent. Fitting the expression in eqn. (3.1) to the results (Huang and Chandler, 2000) from LCW theory yields $\alpha_s = 3.0$ and $T_{0,s} = 0.41$ for the surface term and $\alpha_h = 7.0$ and $T_{0,h} = 0.49$ for the volume solvation term. This assumes that, for the temperature dependence, all amino acids have the same size, while in practice, the volume of amino acids can vary from 75 to 240 Å(Mishra and

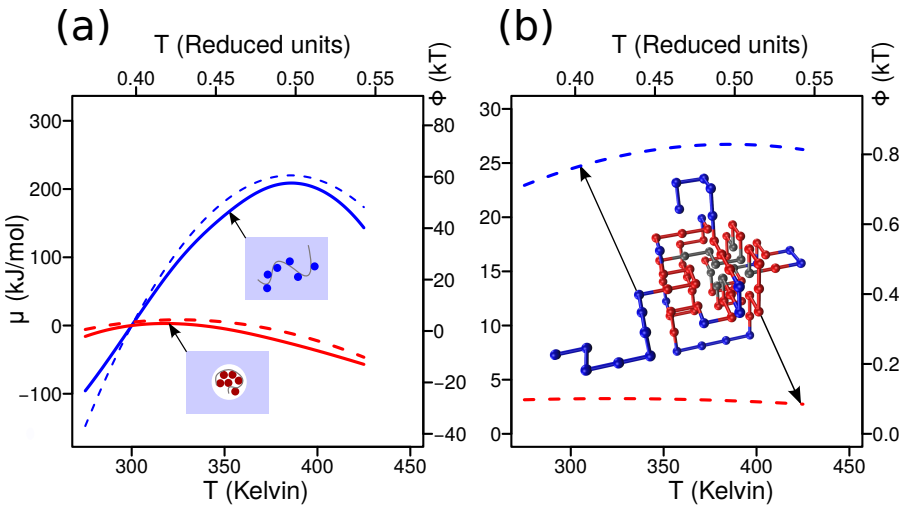

FIGURE 3.1: Comparison between lattice model and LCW-theory for a poly-phenylalanine hydrophobic chain. (a) The chemical potential for a fully extended chain as a function of temperature (blue lines), and the chemical potential for a compacted chain, which we approximate as a 10 Å sphere (red lines). The dashed lines indicate the approximation made by our lattice model, while the solid lines indicate the theoretical predictions from LCW-theory (Lum, Chandler, and Weeks, 1999; Huang and Chandler, 2000). (b) The distinction between surface and fully solvated residues in our model. The blue line shows the potential for the fully solvated residues (corresponding to the residues colored blue), and the red line shows the surface potential (corresponding to the residues colored red).

Ahluwalia, 1984). To test the sensitivity of our model to this assumption, we performed simulations with three different potentials: a temperature independent potential ($\alpha_s = \alpha_h = 0$, a temperature dependent potential (parameters given above) and a strongly temperature dependent potential, corresponding to amino acids that are 15% larger ($\alpha_s = 4.5$ and $\alpha_h = 11.5$) (Derivation shown in SI sections "Derivation temperature dependent potential" and "Approximation of hydrophobicity parameters") (van Dijk et al., 2015).

First, we probe the folding specificity of this model. The lattice model we use here is sequence dependent. In other words, random sequences will typically not fold into a stable structure, whereas designed sequences do so with a high specificity (Shakhnovich et al., 1991; Shakhnovich and Gutin, 1993b; Shakhnovich and Gutin, 1993a; Shakhnovich, 1994; Coluzza, Muller, and Frenkel, 2003; Abeln and Frenkel, 2008; Abeln and Frenkel, 2011).

The number of native contacts (N_{int}) is used as an order parameter for the specificity of protein folding. We define a protein to be folded when $N_{\text{int}} > 75$. The fraction of the simulation spent in this folded state is defined as P_{Fold}. For this work umbrella sampling (Grossfield, 2003) alone is sufficient to sample the configurational space of interest. Figure 3.5S(a) shows that for all potentials the protein folds ($P_{\text{Fold}} > 0.5$) at intermediate temperatures and denatures ($P_{\text{Fold}} < 0.5$) at high temperatures. This is consistent with the view of the high-entropy denatured state caused by the chain entropy. For a well-designed protein, the stability of the protein simulated with a temperature-independent potential is a strictly decreasing function of the temperature, since the native state is optimized to be the lowest enthalpy state.

Only the strongly temperature-dependent potential reproduces cold denaturation as well as heat-induced denaturation, see black curve in Figure 3.5S(a). A very similar folding curve has been observed experimentally for a mutant of cold shock protein Csp (Szyperski et al., 2006). The simulated configurational ensemble of the folded state also includes a small fraction denatured states ($0.20 < T < 0.37$) as observed in the experiment. Note that Csp, like most proteins, does not show cold denaturation above the freezing point of water. However, statistical investigation has shown that the temperature has a measurable influence on the propensity of hydrophobic amino acids to be buried (van van Dijk, Hoogeveen, and Abeln, 2015). This is similar to our observation that proteins become less stable at lower temperatures, but do not denature, for a lower value of the temperature dependence. (Figure 3.5S, green line).

The structural characteristics of the model were investigated by exploring the free energy landscapes of native contacts (N_{int}) and internal contacts between residues (C_{int}); the latter are used as a measure of compactness. At T=0.375, slightly below the transition temperature (T=0.42), two distinct states can be observed, one where the protein is specifically folded ($N_{\text{int}} > 75$), and one in which the protein is mostly unstructured ($N_{\text{int}} < 25$), with a clear barrier separating the two states (Figure 3.5S(c)). Note that the sequence has been designed to fold in this exact structure with 97 native contacts (see Methods).

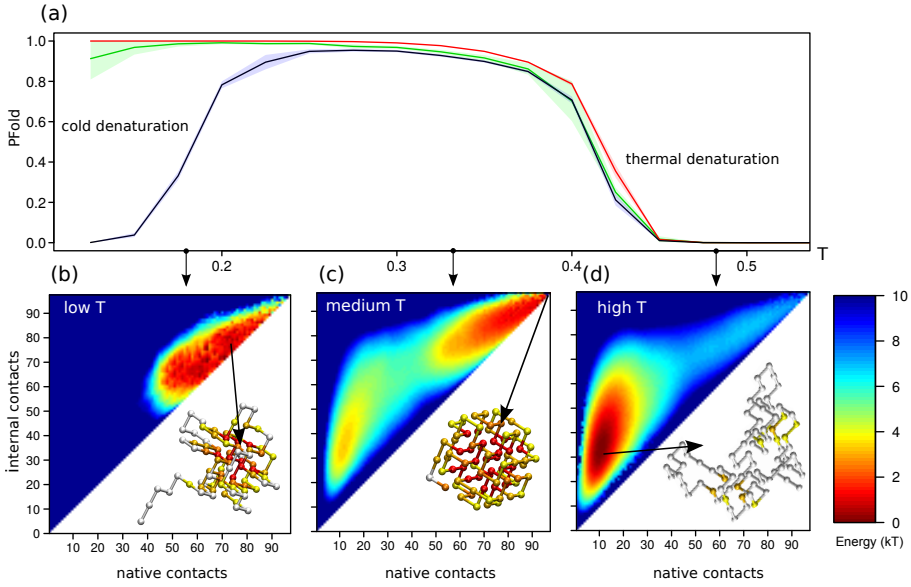

FIGURE 3.2: Temperature-dependent folding stability and structure. The folded state has 97 native contacts. (a) The probability for the model protein to be in the folded state versus temperature, with $\alpha = 0$ (red), the temperature dependent potential (green) and a simulation where the temperature dependence is multiplied by 1.5 (black). The 95% confidence interval is indicated by the shaded area. (b)–(d) Free energy landscapes for the number of native contacts (N_{int}) and all internal contacts (C_{int}) for the simulations with a strong temperature dependence at (b) low temperature ($T = 0.175$), (c) intermediate temperature ($T = 0.375$), and (d) high temperature ($T = 0.475$). For the strong temperature dependent potential, the protein denatures at low temperatures, with many exposed hydrophobic amino acids. However, this denatured structure is a lot more compact than the heat denatured protein, and there are less native contacts present. At intermediate temperatures the protein has the highest stability in its folded configuration (indicated by the arrow) where $N_{\mathrm{int}} = C_{\mathrm{int}} = 97$. At high temperatures the protein makes only transient contacts.

Comparing Figures 3.5S(b) and (d) it becomes apparent that the cold denatured state has more residual structure than the heat denatured state. The cold denatured configurational ensemble at T=0.125 shows a structure that is compact with approximately two thirds of the native contacts present, similar to experimental NMR observations of pressure-assisted cold denaturation (Vajpai et al., 2013), urea-assisted cold denaturation (Wong, Freund, and Fersht, 1996) and cold denaturation for a protein that was destabilized by a mutation (Shan et al., 2010). Note that for some disordered proteins the radius of gyration decreases as the temperature increases (Wuttke et al., 2014; Privalov and Makhatadze, 1993). This is most likely due to interactions involving charged residues, which play a larger role in disordered proteins. Notably, the λ-repressor, which is the most hydrophobic protein in the dataset investigated in Ref. (Wuttke et al., 2014), does show a re-expansion at temperatures higher than 319 K (Wuttke et al., 2014).

In addition to the structural characteristics, our model allows us to investigate the role of the hydrophobic effect in the thermodynamics of protein folding. We start by investigating the heat capacity of folding. Note that the simulations are performed at constant volume, while the experiments are done at constant pressure. However, the difference is neglible for the system in consideration due to the low compressibility of water (Daniel V. Schroeder, 2000) (See SI (van Dijk et al., 2015) for more detail). In order to calculate the heat capacity (C_V in our model) for the temperature-dependent potential, we need to separate the expected enthalpy $< E >$ from the entropic part of the hydrophobic potential, F_{hydr} (see SI (van Dijk et al., 2015)). In a finite system, a phase transition is usually characterized by a sharp peak in the heat capacity that can be observed experimentally (Privalov et al., 1989). For the temperature-independent potential we observe only a single peak at the folding transition (Figure S3(a) (van Dijk et al., 2015)). In contrast, the heat capacity of the temperature dependent potential shows two peaks, one for cold induced denaturation and one for heat induced denaturation (Figure S3(b) (van Dijk et al., 2015)). Another interesting observation is a linear temperature dependence of the heat capacity in the temperature range where no phase transition is occurring (Figure 3.6S(a)). The slope or the linear increase in the baseline of the heat capacity has been investigated in Refs. (Privalov and Dragan, 2007; Muñoz and Sanchez-Ruiz, 2004; Farber et al., 2010; Bruscolini and Naganathan, 2011; Johnson, 2013). In the context of the current model we can understand the linear T-dependence of the heat capacity in terms of the exposed hydrophobic groups. Assuming a constant hydrophobe-water contact area and neglecting entropic contributions other than the hydrophobe-water contact area, we can derive a simple lower bound for the heat capacity (see SI (van Dijk et al., 2015) for derivation):

$$C_V(T) = (2\alpha_s N_{\mathrm{s}} + 2\alpha_h N_{\mathrm{h}})T \tag{3.2}$$

where N_{s} is the number of hydrophobic amino acids that are at the surface, and N_{h} the number of hydrophobic amino acids that are fully hydrated. This

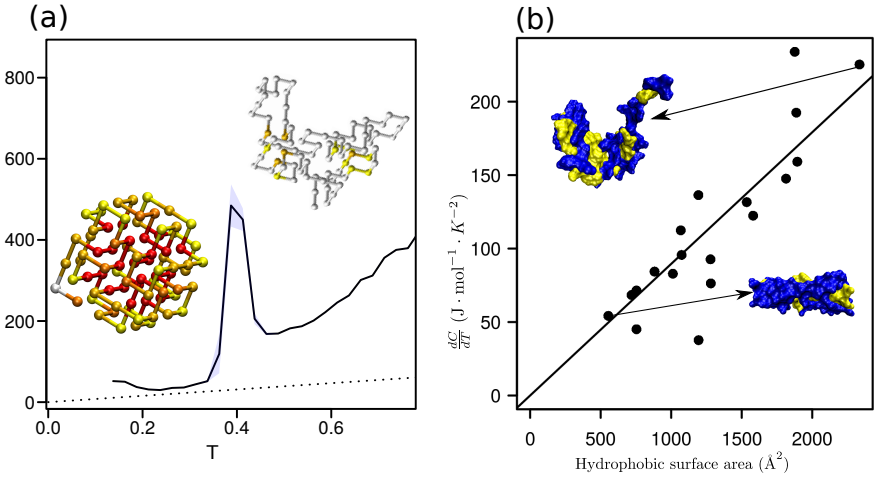

FIGURE 3.3: Heat capacity versus temperature (a) The heat capacity for the fitted potential as calculated by $C_V = \frac{dE}{dT}$, corresponding to the green line in Figure 3.5S(a), shows a linear increase in the heat capacity with respect to temperature. Our model suggests that the slope of this baseline, which is a lower bound for the heat capacity, should correlate with the amount of hydrophobic surface area. (b) The slope of the heat capacity $\frac{dC}{dT}$, shows a strong correlation with the exposed hydrophobic surface area in real protein structure as predicted by our model. Two indicative protein structures, with pdb-codes 1J46 and 2ZTA are shown with the hydrophobic and hydrophilic amino acids colored yellow and blue respectively.

simple calculation yields a lower bound for the heat capacity in regions where no folding transition occurs, as indicated by the dotted lines in Figure 3.6S(a). Here we estimate N_s and N_h for the folded regime from the number of exposed hydrophobes in the native structure ($N_\mathrm{s} = 13$, $N_\mathrm{h} = 0$). The difference between the lower bound shown in Figure 3.6S(a) and the simulated results is likely due to the chain entropy in the native ensemble. This is supported by the results from the temperature independent potential (see also Figure S3(a) (van Dijk et al., 2015)).

Equation (3.2) has additional consequences that can be verified experimentally. Initially, as a consistency check, it is easy to see that from eqn. (3.2) we can recover the well known relationship between the change in heat capacity, ΔC_V and the change in hydrophobic surface area upon folding at a given temperature T, eg. Refs (Spolar, Ha, and Record, 1989; Anslyn and Dougherty, 2006). We can go further, however, and probe the derivative of the heat capacity with respect to the temperature ($\frac{dC_V}{dT}$), as shown by the dotted lines in Figure 3.6S(a). Our analysis in eqn. (3.2) predicts that this slope itself is proportional to the exposed hydrophobic surface area (corresponding to N_h in the model) in a given state.

To test this prediction we compared the heat capacity slopes for folded proteins tabulated in Ref. (Privalov and Dragan, 2007) with the level of exposed hydrophobic surface area in the corresponding folded structures obtained with (Kabsch and Sander, 1983). We find a strong correlation ($R^2 = 0.77$) between the slope of the heat capacity and the exposed hydrophobic surface area (see Figure 3.6S(b)), further supporting the temperature dependence of the hydrophobic effect as a key mechanism underlying the linear increase of the heat capacity.

Previously, a higher slope for DNA-binding proteins has been observed when compared to globular proteins of the same size Gómez et al., 1995; Privalov and Dragan, 2007; Naganathan et al., 2011a; van Dijk et al., 2015, see also SI Figure 2 was rationalized through the flexibility of DNA-binding domains (Privalov and Dragan, 2007; Naganathan et al., 2011a). Our work suggests that the increased slope of the heat capacity may be explained solely by a higher exposed hydrophobic surface area. Hence, the strong correlation between the slope of the heat capacity and the exposed hydrophobic surface area for all proteins may be explained by a consistent treatment of hydrophobicity alone (see Figure S2 (van Dijk et al., 2015)). This does not, however, preclude a possible correlation between the flexibility and the amount of exposed surface area.

To conclude, we have presented an extension for a coarse-grained lattice model by including a temperature-dependent hydrophobic term in the interaction potential. The combination of the coarse-grained steric model and the potential ensures appropriate contributions of the solvent-residue and internal contacts. This allows us to separate chain entropy from the solvent entropy. An effective quadratic potential is applied to account for the temperature dependence of the hydrophobic effect. Simulating the model, we observe cold

denaturation for realistic parameter settings, suggesting that the hydrophobic effect is the key component in cold denaturation. In addition, we find that the simulated cold denatured state is more compact than the heat denatured state; this is in agreement with experimental observations. Moreover, the model is able to reproduce the characteristic experimental heat capacity curves for protein folding.

Starting from the temperature dependent potential, we derive a simple relation that approximates the heat capacity baseline of the native state of the protein. This relation is tested for a set of real proteins, where we do indeed find a correlation between the hydrophobic surface area and the slope of the heat capacity. Hence, our model seems to make accurate predictions for thermodynamic behaviour of real proteins. Additionally, the developed relation can potentially be used to calculate an accurate baseline for proteins with a known structure.

The interaction potential is based on a representative set of the Protein Database, PDB-25 (Griep and Hobohm, 2010), and holds some biases towards rigid and soluble proteins, since they are easier to crystallize (see SI (van Dijk et al., 2015) for further discussion). The cubic lattice model also has some limitations: secondary structure can not be modelled explicitly, nor is there sufficient molecular detail to predict the true fold of a protein sequence. However, we stress that the use of our coarse-grained model is not only motivated by considerations of computational cost. Rather, the use of simple, coarse-grained models allows us to reveal the minimal physical ingredients that a model needs to account for cold denaturation.

The temperature-dependent potential developed here is applicable to other (off-lattice) coarse-grained models with implicit solvent-side chain interactions, e.g. Refs. (Hoang et al., 2004; Auer et al., 2008; Coluzza, 2011). The protocol to calculate the heat capacity from a temperature-dependent potential, can be applied to any effective potential with a closed form expression that is continuous and differentiable with respect to $\beta = \frac{1}{k_B T}$. Furthermore, the addition to the lattice model itself will enable investigation of temperature dependence of protein aggregation building on previous studies (Abeln and Frenkel, 2008; Ni et al., 2013b; Abeln et al., 2014b)

3.1 Supplementary Material – Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation

Monte Carlo simulations

The simulations were run using the Metropolis-Monte Carlo algorithm. This method consists of making a random trial move, evaluating the Hamiltonian and accepting using the Boltzmann criterium:

$$P_{Acc} = \max(1, \exp(-\mathcal{H}/k_B T)) \tag{3.3}$$

where $\mathcal{H}$ is the Hamiltonian, k_B is the Boltzmann constant, and T is the temperature. The complete Hamiltonian of the system is obtained as the sum of the base pair potential and the temperature-dependent correction F_{hydr} and (defined in eqns. (3.5) and (3.6) respectively):

$$\mathcal{H}(T, \vec{r}) = E_{\mathrm{base}} + F_{\mathrm{hydr}} \tag{3.4}$$

We calculate the interactions with the solvent separately for surface terms and fully solvated residues

For E_{base}, we use a standard cubic lattice model, first introduced in Ref. (Sali, Shakhnovich, and Karplus, 1994) (Figure 1). A more detailed description of the model used here can be found in Ref. (Abeln and Frenkel, 2011). In this model, the potential energy E_{base} of the amino acid interactions can be calculated as follows:

$$E_{\mathrm{base}} = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\epsilon_{a(i),a(j)} \cdot C_{ij} + \sum_{i}^{N}\epsilon_{a(i),w}C_{\mathrm{wi}} + \sum_{i}^{N}\omega_s\epsilon_{a(i),w}C_{\mathrm{si}} \tag{3.5}$$

Here, N is the number of amino acids, $\epsilon_{a(i),a(j)}$ represents the interaction of amino acid $a(i)$ and $a(j)$. C_{ij} is an adjacency matrix, and is 1 if residues i and j are neighbours in the lattice. A residue can be categorized as buried, ($C_{\mathrm{wi}} = 0$ and $C_{\mathrm{si}} = 0$), on the surface ($C_{\mathrm{wi}} = 0$ and $C_{\mathrm{si}} = 1$), or fully solvated ($C_{\mathrm{wi}} = 1$ and $C_{\mathrm{si}} = 0$). We define a residue to be buried if it has zero contacts with water, on the surface if it has 1-3 contacts with water, and fully solvated if it has 4 or more contacts with water. The weigth ω_s accounts for the lower interaction energy for a hydrophobic residue at the surface.

Derivation temperature dependent potential

The MJ matrix represents effective interaction free energies. This means it should contain temperature-dependent effects. While van der Waals, electrostatic and other interactions are mostly enthalpic, the hydrophobic interactions are partly entropic and partly enthalpic in nature. Thus, solvent exposed hydrophobic amino acids exhibit a temperature dependence in their excess chemical potential of solvation which is characteristic of small hydrophobes. For small solutes with a radius $r \lesssim 0.5$nm, their free energy of solvation is proportional to the volume of the solutes and they can be accommodated without breaking of hydrogen bonds in the water. In this scenario, when the hydrophobic amino acids in the model protein are fully solvated, N_h is equal to the total number of hydrophobic amino acids. When the hydrophobic amino acids cluster at the core of the protein as it folds, N_s will describe the number of surface exposed amino acids and therefore in this regime the solvation energy emerges as a surface energy as predicted by theory (Weeks, Chandler, and Andersen, 1971; Chandler, 2005). Indeed, in this regime it is impossible to accommodate the cluster of hydrophobes while maintaining a complete hydrogen bonding network in the surrounding water and hence there is a driving force for water molecules to vacate the hydrophobic cluster and an interface is formed between the bulk water and the water-depleted interior of the model protein. The energy cost for solvating the cluster of hydrophobes is therefore dependent on the surface rather than the volume of the solute, as well as on the surface tension of the air/water interface.

This means that to correctly model the free energies as a function of temperature, a correction needs to be applied to the interaction columns. For the temperature-dependent potential developed here, we use a quadratic potential:

$$F_{\mathrm{hydr}} = -\alpha_s N_{\mathrm{s}}(T - T_{0,s})^2 - \alpha_h N_{\mathrm{h}}(T - T_{0,h})^2 \tag{3.6}$$

T_0 and α are constants. The subscript s and h indicate the surface and fully hydrated terms. The complete Hamiltonian of the system is obtained as the sum of the temperature-dependent correction F_{hydr} and the base pair potential (defined in eqns. (3.4), (3.5) and (3.6)) respectively.

Approximation of hydrophobicity parameters

To relate our model to theoretical calculations, we introduce a new function, $\phi(T)$, which corresponds to chemical potential of solvation for a single residue. For a single fully hydrated residue, $\phi_h(T) = \mu$. This means that for any single configuration of a lattice protein, we can calculate the chemical potential of solvation as $\mu_{\mathrm{lattice}} = N_s\phi_s(T) + N_h\phi_h(T)$.

We calculate $\phi(T)$ separately for surface terms and fully solvated residues. For a fully solvated residue, the interaction is approximated by a quadratic fit to the potential for a 3.3 Å particle to data in Ref. (Huang and Chandler, 2000). Since the model is run in reduced units, this potential needs to be rescaled

to be consistent with the statistical potential. T_{amb}, is taken to be 300K, where proteins are stable and their structure is being determined. $\Phi_h(T_{\mathrm{amb}})$ is the corresponding interaction of the hydrophobic particle with water in the statistical potential. This yields a reference point for the potential, allowing the potential to be rescaled to the interactions in the lattice model. We write the reduced temperature, T^*, as: $T^* = kT$. In our model, $T^*_{\mathrm{amb}} = 0.4$, which corresponds to 300 K. Solving for k yields $k = 1.3 \cdot 10^{-3} K^{-1}$. Unless otherwise specified, temperatures in this paper are given in reduced units.

The weight of the surface term, ω_s, used to rescale the surface potential using:

$$\Phi_s(T_{\mathrm{Amb}}) = \omega_s \Phi_h(T_{\mathrm{amb}}) \tag{3.7}$$

is fitted by considering a fully hydrophobic chain in the native structure of the lattice model, counting the number of surface terms, and distributing the solvation free energy over the surface in our model. The native structure consists of 24 buried residues, 1 fully solvated residue and 55 surface residues. We use a sphere with a radius of 10 Å to approximate the size of a protein.

The ratio of the surface terms with respect to the volume terms is fitted such that the ratio of the transfer free energy, μ, of the 10 Å particle to the 3.3 Å is the same as in the theoretical model. The values resulting from this procedure: $\alpha_s = 3.0$, $T_0 = 0.41$ and $\omega_s = 0.41$ for the surface terms and $\alpha_h = 7.0$ and $T_0 = 0.49$ for the fully solvated terms.

$$\frac{\mu_{\mathrm{LCW}}(r_{\mathrm{protein}}, T_{amb})}{\mu_{\mathrm{LCW}}(r_{\mathrm{residue}}, T_{amb})} = \frac{N_s \omega_s \phi_{s,\mathrm{PHE}}(T_{\mathrm{amb}}) + N_h \phi_{\mathrm{h,PHE}}(T_{\mathrm{amb}})}{\phi_{\mathrm{s,PHE}}(T_{\mathrm{amb}})} \tag{3.8}$$

The subscript PHE indicates that we are setting the interaction of Phenylalanine amino acid with water at ambient temperature, $\epsilon_{\mathrm{PHE},w}$. Now, $\phi_{\mathrm{h,PHE}}(T_{\mathrm{amb}}) = \epsilon_{\mathrm{PHE},w}$. Since μ_{LCW} can be obtained from theory (Huang and Chandler, 2000), we can now obtain ω_s by combining eqns. (3.7) and (3.8).

We compare the results of the fitted parameters to the case $\alpha_s = \alpha_h = 0$. Here, the Hamiltonian corresponds to the temperature-independent potential fitted in Ref. (Abeln and Frenkel, 2011) with the adjustment that $\omega_s = 1$ was used there.

Real amino acids vary in volume from 72 to 240 $\mathrm{Å}^3$ (Mishra and Ahluwalia, 1984). The above derivation assumes amino acids with a 3.3Å radius, corresponding to a volume of 150 $\mathrm{Å}^3$, We also derive the potential for amino acids with a size of 225 $\mathrm{Å}^3$, corresponding to a radius that is 15% larger. This yields the following values: $\alpha_s = 4.5$ and $\alpha_h = 11.5$.

Definition of folded state

Since in our simulations the folded ensemble is somewhat fluid and can contain configurations which are close to native, we need a way to define the folded

ensemble. For this, we determined the point where the free energy with respect to the number of native contacts becomes significantly higher than zero. This was based on multiple free energy profiles. A representative free energy profile is shown in Figure 3.4S.

Heat Capacity of the system

Validation of the model is done by determining the heat capacity of the system through its definition, $C_V = \frac{dQ}{dT}$, which can be evaluated in this model through

$$C_V = \frac{d\langle E\rangle}{dT}, \tag{3.9}$$

provided that no work is done on the system. In order to calculate $\langle E\rangle$ we need to separate the hydrophobic potential, F_{hydr}, into an enthalpic and entropic contribution. This can be calculated using eqn. (3.6) and the identity:

$$\langle E\rangle = \frac{d\beta F}{d\beta} \tag{3.10}$$

Note that $\langle E\rangle = \langle E_{\mathrm{base}}\rangle$ for a temperature-independent potential. For the temperature-dependent potential this expression evaluates to $\langle E\rangle = \langle E_{\mathrm{int}}\rangle + \langle E_{\mathrm{hydr}}\rangle$. At a fixed number of hydrophobe-water contacts we can calculate $\langle E_{\mathrm{hydr}}\rangle$ analytically.

Since F_{hydr} is the only temperature-dependent part of the Hamiltonian $\mathcal{H}(T,\vec{r})$ defined in eqn. (3.4), $\langle E\rangle = \langle E_{\mathrm{base}}\rangle + \langle E_{\mathrm{hydr}}\rangle$, which can then be used to numerically estimate the derivative $\frac{d\langle E\rangle}{dT}$. The results are shown as black lines in Figure 3.

Comparison heat capacity to experiments

The heat capacity in the simulations are calculated at a constant volume (C_V), while the heat capacity is generally measured at constant pressure in experiments (C_P). While for gases the heat capacity can vary significantly, for water this difference is generally negligible. This can be shown by considering the identity for the difference in heat capacity at constant pressure and constant volume (Daniel V. Schroeder, 2000):

$$C_P - C_V = VT\frac{\alpha^2}{\beta_T} \tag{3.11}$$

Where α is the thermal expansion coefficient, and B_T is the isothermal compressibility, V is the volume, and T is the the temperature in K. For water, $\alpha = 69{\cdot}10^{-6}\mathrm{K}^{-1}$, and $\beta_T = 4.6{\cdot}10^{-10}\mathrm{m}^2\mathrm{N}^{-1}$. For water the difference between the heat capacity at constant pressure and the heat capacity at constant volume is approximately 0.6 % at 293 K. The presence of a diluted protein does not significantly affect the isothermal compressibility and the thermal expansion

coefficient. Therefore, the difference between the heat capacity at constant volume and the heat capacity at constant pressure is not significantly affected by the presence of diluted protein, which means that that we can directly compare the experimental and simulated heat capacities.

Obtaining hydrophobic surface area

We use the DSSP program (Kabsch and Sander, 1983) to determine the surface accessible area of each residue for PDB-structures. We then define the following amino acids to be hydrophobic, in accordance with the amino acids that are considered hydrophobic in the lattice model: Alanine, phenylalanine, cysteine, leucine, isoleucine, tryptophan, valine, methione, and tyrosine. The sum of these contributions is shown on the y-axis in Figure 3C. The slope of the heat capacity is taken from (Privalov and Dragan, 2007) and multiplied by the sequence length found in the corresponding Protein-Database structure.

Design procedure

Because of the level of coarse graining present in the model, and the simplifying assumptions made in the potential, existing proteins will not fold to a correct structure. To find a sequence that will fold into a structure, it has to be specifically designed. The design procedure introduced by Shakhnovich and Gutin (Shakhnovich and Gutin, 1993b; Shakhnovich and Gutin, 1993a; Shakhnovich, 1994) and used with small adaptations in Refs. (Coluzza, Muller, and Frenkel, 2003; Abeln and Frenkel, 2008; Abeln and Frenkel, 2011) uses a combination of the minimization of the potential energy of the folded structure and a measure to keep the variance of the amino acids high using a Monte Carlo algorithm:

$$P_{\mathrm{Acc1}} = \min(1, \exp(-\Delta E_{\mathrm{base}} \beta_2)) \tag{3.12}$$

where P_{Acc1} is the energy acceptance probability, E_{base} is the base pair potential energy of the folded state and β_2 is 1 divided by the design temperature.

We use a slightly adapted method where the potential energy difference between the native state and the fully extended state is minimized. This yields structures with a slightly higher heat denaturation temperature and a larger contribution of the hydrophobic interactions to the stability.

Protein sequence

The protein sequence containing 80 amino acids was designed for a structure with 97 native contacts, and has a clear hydrophobic core.

The sequence is given as:

```
PVVNL TSPEF FKSKL WGDVN WMRQT PEIHM CLQHK
QNHRQ GEMCY GYCAP SCLEH ECGRN DDYRM SKFID
YVKIF ATWTA
```

Discussion of pair potential

The pair-potential in our paper, $\epsilon_{a(i),a(j)}$ is calculated with the same procedure as Hobohm & Sander, but uses a larger dataset (see Ref (Abeln and Frenkel, 2011)) based on PDB-select-25 (Griep and Hobohm, 2010). Due to the discrete nature of the lattice model long range interactions are not properly accounted for in the model. Since charged interactions are important for flexible and natively disordered proteins, this potential should not be used to model these proteins. Moreover, since flexible structures are harder to crystallise, the data contains a bias to soluble, rigid proteins. Nonetheless, a comparison potential derived from NMR-structures, where this bias is less pronounced, shows no significant differences in amino acid interactions derived from X-ray structures (van Dijk, Hoogeveen, and Abeln, 2015).

The temperature dependence of the free energy of transfer from the core of the protein to the solvent for hydrophobic residues is derived from hydrophobic theory. In previous work studying temperature dependent effects, this temperature dependence has been estimated from laboratory measurements of transfer (Privalov and Makhatadze, 1993) from a gaseous environment to water (Wuttke et al., 2014). While this approach provides a separate temperature dependence for each amino acid, the gas state may not capture the internal environment of the folded protein accurately. Nonetheless, for the hydrophobic residues this procedure yields values that are qualitatively similar to the results we obtain by using a statistical pair potential as a reference point and adding a temperature dependent term to the potential. A third approach, deriving a statistical pair potential from a large set of NMR-structures determined at differing temperatures, also yields qualitatively similar results. The quantitative differences between the potentials derived by the three methods can be explained by the differences in methodology.

The potential used is included in the following file:

Potential_Data.txt

The data represents a two-dimensional array, where the order of the columns is the same as the order in the rows.

Umbrella sampling

This simple model allows for sufficient sampling of the conformational space at high temperatures with respect to the number of native contacts using simple Monte Carlo sampling. Umbrella sampling with the following quadratic biasing potential was performed to verify the results obtained in this fashion:

$$E_{\text{umbr}} = \mathcal{H}(T, \vec{r}) + k(N_{\text{int}} - N_0)^2, \tag{3.13}$$

where $\mathcal{H}(T, \vec{r})$ is the normal Hamiltonian, N_{int} is the number of native contacts k is the spring constant, and N_0 is the centre of the biasing potential. The

spring constant k was set to 0.02, and N_0 was set to values between 0 and 100, with a stepsize of 5, so $N_0 \in \{0, 5, 10, 15, ..., 95, 100\}$. WHAM (Grossfield, 2003) was then used to reconstruct the free energy surface with respect to N_{int}. We define a protein to be in the folded state when $N_{\text{int}} > 75$. The maximum number of native contacts is 97.

To explore the landscape of the native contacts together with the internal contacts C_{int}, umbrella sampling in two dimensions was performed. The potential used in this case was:

$$E_{\text{umbr}} = \mathcal{H}(T, \vec{r}) + k((N_{\text{int}} - N_0)^2 + (C_{\text{int}} - C_0)^2), \tag{3.14}$$

where all values were identical to the previous simulations. The simulations were run with values for C_0 ranging from 0 to 100 internal contacts, with an interval of 5, so $C_0 \in \{0, 5, 10, 15, ..., 95, 100\}$. Again, WHAM was used to reconstruct the free energy surface with respect to C_{int}. Note that while it is possible to have more than 97 internal contacts, this was not encountered often in our simulations and those configurations therefore had a high free energy.

Derivation Heat Capacity

The heat capacity is given by $C_V = \frac{d\langle E\rangle}{dT}$ where $\langle E\rangle = \frac{d\beta F}{d\beta}$ is the expected value of the internal energy of the system. For a given configuration, the free energy F equals the Hamiltonian $\mathcal{H}(T, \vec{r})$. We can use this to find we can find the expected potential energy of the system, $\langle E\rangle$ for a configuration.

$$F = E_{\text{base}} + F_{\text{hydr}} \tag{3.15}$$

Substituting F_{hydr} using eqn. (3.6)

$$F = E_{\text{base}} - \alpha_s N_{\text{s}}(T - T_{0,s})^2 - \alpha_h N_{\text{h}}(T - T_{0,h})^2 \tag{3.16}$$

Multiply both sides by β.

$$\begin{aligned}\beta F = \beta E_{\text{base}} - \alpha_s N_s(\frac{1}{\beta} - 2T_{0,s} + \beta T_{0,s}^2) - \\ \alpha_h N_h(\frac{1}{\beta} - 2T_{0,h} + \beta T_{0,h}^2)\end{aligned} \tag{3.17}$$

Taking derivative with respect to β and using that $\frac{d\beta F}{d\beta} = \langle E\rangle$

$$\langle E\rangle = E_{\text{base}} - \alpha_s N_s(T_{0,s}^2 - T^2) - \alpha_h N_h(T_{0,h}^2 - T^2) \tag{3.18}$$

N_{s} represents the number of residues on the surface, while N_{h} represents the number of residues that are fully solvated. The average over the configurational ensemble can subsequently be determined through Monte Carlo sampling, and the heat capacity can be determined through a numerical evaluation of eqn. (3.9).

Assuming the folded structure consists of a single configuration, the heat capacity of the folded structure can be determined analytically:

$$C_V(T) = \frac{d\langle E\rangle}{dT} = 2\alpha_s N_s T + 2\alpha_h N_h T \tag{3.19}$$

Length scale dependence of solvation

The excess chemical potential of solvation, μ, shown in Figure 3.1S is calculated by considering a hydrophobic homopolymer configured in a cubic formation of increasing dimensions. If the temperature is fixed, μ can be written as the sum of the residue-solvent interactions:

$$\mu = r_s\epsilon_{s,w} + r_h\epsilon_{h,w} \tag{3.20}$$

Where r_s the number of residues at the surface, $\epsilon_{s,w}$ is the interaction energy for a surface residue with the solvent, and $\epsilon_{h,w}$ is the interaction of a fully solvated residue with the solvent. In our model, $\epsilon_{s,w} = \omega_s\epsilon_h$. As described in the methods, the weight ω_s is fitted to 0.32. Note that buried residues do not contribute to the transfer chemical potential since they are not in contact with the solvent.

To compare the relation of the chemical potential with the size of the hydrophobic cluster in our model, we consider a cube of different dimensions. Let r_e be the number of residues along an edge. Except for the special case $r_e = 1$, there are no fully solvated residues. That means that the chemical potential is:

$$\mu = r_s\epsilon_{s,w} \tag{3.21}$$

Therefore, computing the number of surface residues allows us to find the chemical potential. This is achieved by subtracting the number of residues that are buried, $(r_e - 2)^3$, from the total number of residues, r_e^3.

$$r_s = r_e^3 - (r_e - 2)^3 = 6r_e^2 - 12r_e + 8 \tag{3.22}$$

This means that the chemical potential can be written as:

$$\mu = \epsilon_{s,w}(6r_e^2 - 12r_e + 8) \tag{3.23}$$

The result of this equation is plotted in Figure 3.1S.

Simplified model

To test the influence of the length scale dependence on the results, we have included simulations with a simplified model that does not incorporate this length scale dependence, by setting $\alpha_s = \alpha_h$. This results in qualitatively similar results, as shown in Figures 3.5S and 3.6S.

Table of proteins

Protein	PDB	$\frac{dC_P}{dT}$	Hydrophobic surface area
BPTI	5pti	37.7	1194
Barnase	1rnb	76.3	1281
Myoglobin	1mbo	122.4	1581
Lysosyme	1lzl	45.1	754
Cytochrome C	5cyt	92.7	1279
Ubiquitin	1ubq	68.4	719
T_4 lysozyme	3lzm	147.6	1814
RNase T1	8rnt	95.68	1074
RNase A	7rsa	136.3	1193
Engrailed	3hdd	71.5	754
Mat α 2	1apl	84.37	882
Antennapedia	9ant	82.88	1012
HGMD-74	1hma	112.42	1068
LZ-GCN4	2zta	54.25	554
HMG SOX5	1I11	131.6	1536
Zn-finger TFIIIA	1tf3	159.16	1894
NHP6A	1cg7	192.51	1887
SRY	1j46	233.92	2335
Lef-79	2lef	225.25	1876

Lists the PDB-ids, the slope of the heat capacity and the hydrophobic surface area for the proteins investigated.

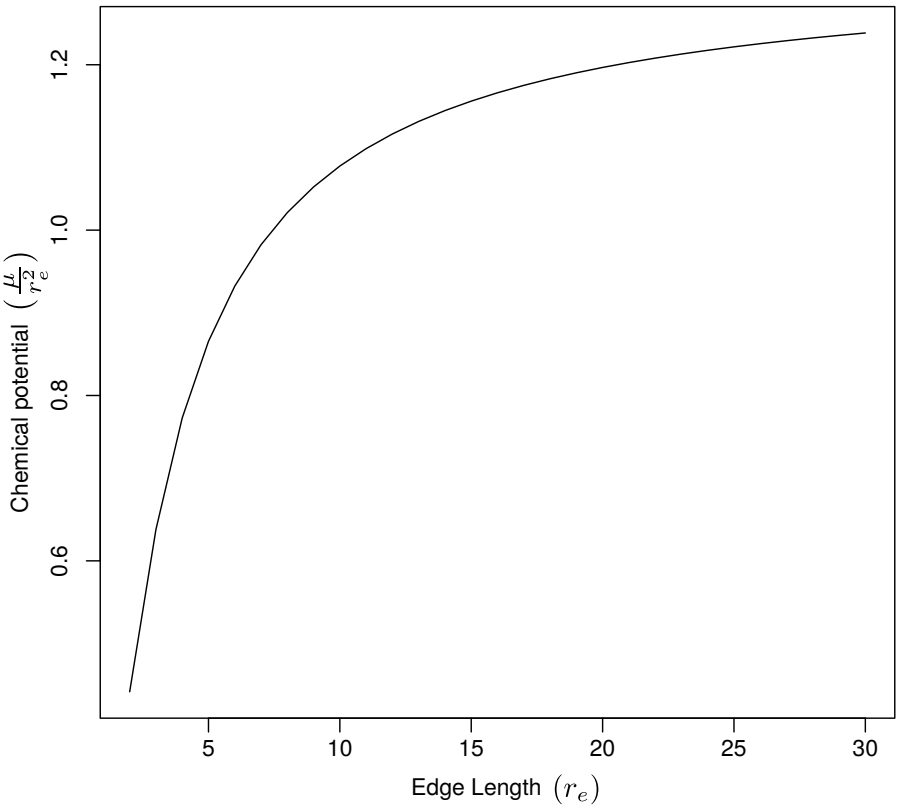

FIGURE 3.1S: The length-scale dependence of the excess chemical potential of solvation, μ, as a function of the edge length of a hydrophobic cube, r_e, in the model. The values of the chemical potential are shown for $2 \leq r_e \leq 30$. Theoretically (Huang and Chandler, 2000; Chandler, 2005), the chemical potential scales with the volume of a solvated hydrophobe for small solutes, and with the surface for large solutes. A similar scaling is found in our model for both small and large solutes.

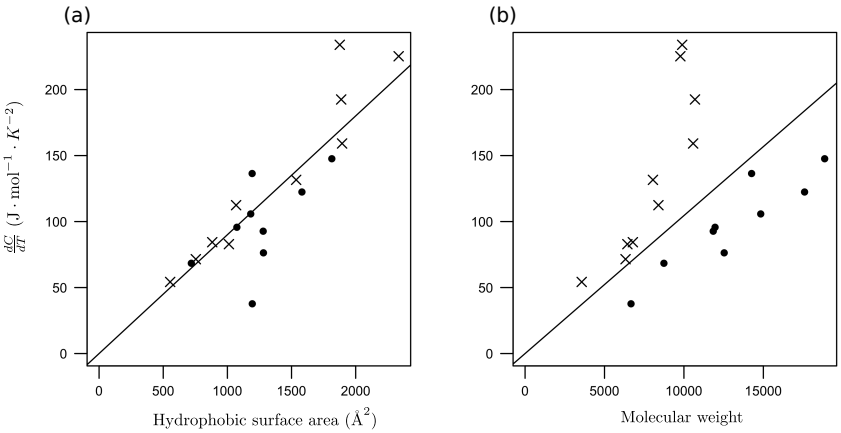

FIGURE 3.2S: Hydrophobic surface area vs Slope of heat capacity (a) and Molecular weight vs Slope of heat capacity (b). A similar analysis in (Naganathan et al., 2011a) for an extended set of proteins, shows the same pattern but does not recognize that the hydrophobic surface area provides a fit that is valid for both normal globular proteins (dots) and DNA-binding proteins (crosses).

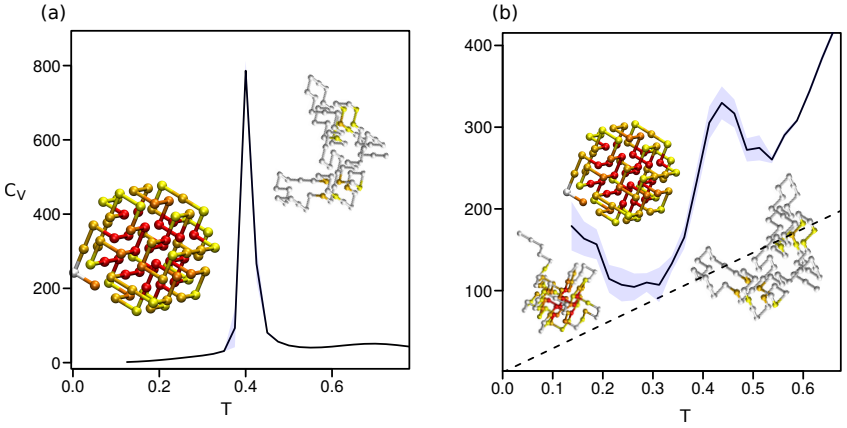

FIGURE 3.3S: Heat capacity for temperature independent potential (a) and temperature dependent potential (b). (a) Heat capacity for a temperature independent potential, similar to earlier published results, we find a single peak for the heat induced denaturation. (b) Heat capacity versus temperature for bigger amino acids. A double peak can be seen for the heat capacity, indicating two phase transitions.

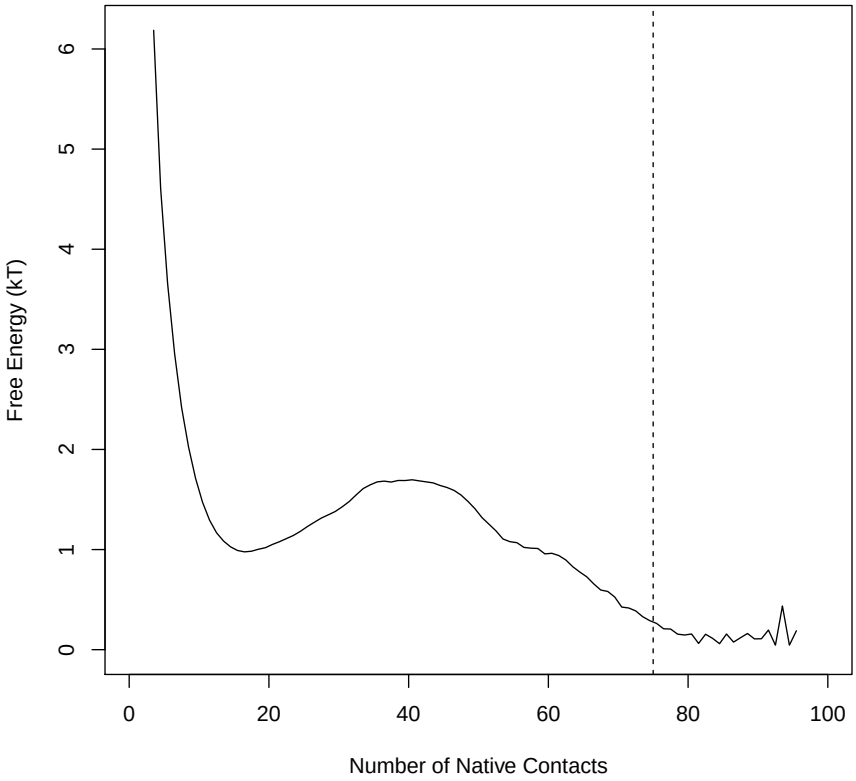

FIGURE 3.4S: Free Energy landscape with respect to number of native contacts for the temperature dependent potential at $T = 0.4$. Since in our simulations the folded ensemble is somewhat fluid and can contain configurations which are close to native, we need a way to define the folded ensemble. Moreover, there are some artifacts near the native state in the energy landscape due to the discrete nature of the lattice model. The definition of the folded state was determined as the point where the free energy with respect to the number of native contacts becomes significantly higher when compared to the minimum.

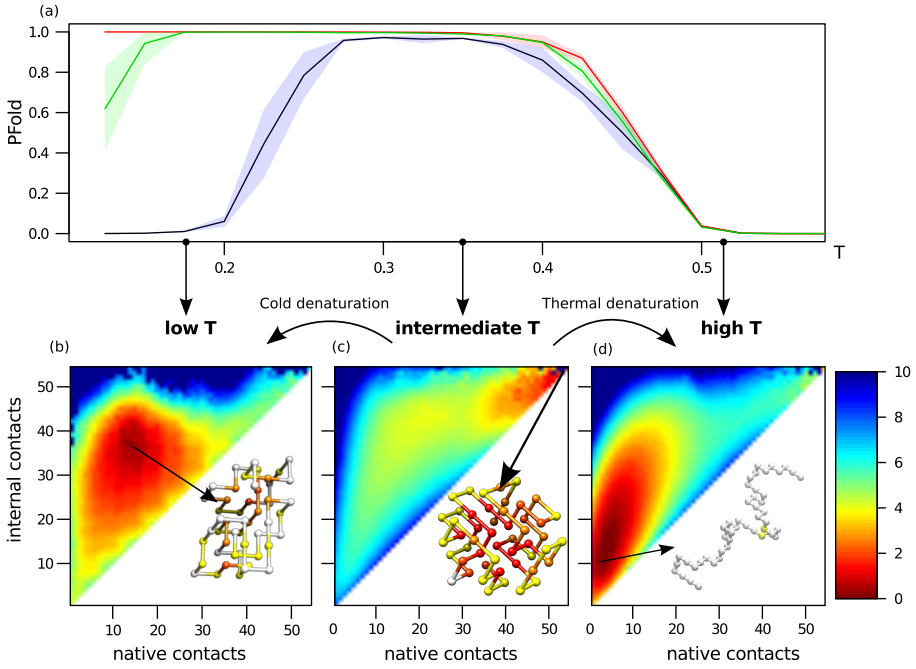

FIGURE 3.5S: Temperature-dependent folding stability and structure.

(a) The probability for the model protein to be in the folded state, P_{Fold}, versus temperature, with $\alpha_s = \alpha_h = 0$ (red), $\alpha_s = \alpha_h = 5k_BT$ (green) and $\alpha_s = \alpha_h = 10k_BT$ (black). A protein is considered folded if more than 45 of the 54 possible native contacts are present. The 95% confidence interval is indicated by the shaded area.

(b)–(d) Free energy landscapes for the number of native contacts (N_{int}) and all internal contacts (C_{int}) for $\alpha_s = \alpha_h = 10k_BT$ at (b) low temperature ($T = 0.175$), (c) intermediate temperature ($T = 0.375$), and (d) high temperature ($T = 0.575$). For $\alpha_s = \alpha_h = 10k_BT$ the protein denatures at low temperatures, with many exposed hydrophobic amino acids. However, this denatured structure is a lot more compact than the heat denatured protein, and there are less native contacts present. At intermediate temperatures the protein has the highest stability in its folded configuration (indicated by the arrow) where $N_{\text{int}} = C_{\text{int}} = 54$. At high temperatures the protein makes only transient contacts.

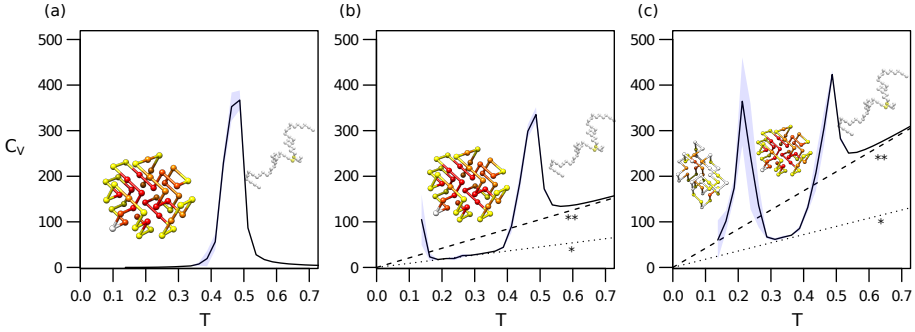

FIGURE 3.6S: Heat capacity versus temperature for $\alpha_s = \alpha_h = 0$ (a), $\alpha_s = \alpha_h = 5k_BT$ (b) and $\alpha_s = \alpha_h = 10k_BT$ (c). 95 % confidence region is shown through shading. The dotted lines indicate the baselines of the native state (*) and the denatured state (**), obtained through a simple calculation (see section "Derivation Heat Capacity").
(a) The classic model shows a single peak in the heat capacity at the folding transition. (b) The heat capacity for $\alpha_s = \alpha_h = 5k_BT$, as calculated by $C_V = \frac{dE}{dT}$ does not show cold denaturation, but does show a linear increase in the heat capacity with respect to temperature. (c) The heat capacity clearly shows two sharp peaks signifying two phase transitions. The linear increase in heat capacity in the absence of a phase transition is also observed experimentally.

Chapter 4

BICEP: Heat Capacity Baseline Prediction

Based on the manuscript:

E van Dijk, R Bouwmeester, BJ Brandt, J Heringa, and S Abeln (2016b). "Heat Capacity Baseline Prediction Using BICEP". in: *In preparation*

Abstract

Heat capacity is a well-defined experimental observable that can be measured using Differential Scanning Calorimetry (DSC). Experimental observations show that native proteins have a linear increase in the baseline of their heat capacity. This baseline can give important clues about the state of the folded protein. Currently the linear part of the experimental baseline can be predicted using the Freire baseline estimate, which is based on the molecular weight of the protein. Here we show that using a combination of the Hydrophobic Surface Area (HSA) and the molecular weight gives a better prediction for the linear part of the experimental baseline. Moreover, this approach for baseline estimation can be rationalized through current hydrophobic theory. We make the baseline estimates available through our web server: the Baseline Increase in heat Capacity Estimation with hydroPhobicity (BICEP `http://www.ibi.vu.nl/programs/bicepwww/`). Two separate use cases show how the output of BICEP may be used to interpret experimental baselines.

Introduction

Differential Scanning Calorimetry (DSC) is a widely used experimental technique to investigate protein-folding transitions by measuring the heat capacity with respect to temperature. Proteins with a two-state folding transition usually show a sharp peak in the heat capacity around their melting temperature. The surface area under the peak can be used to determine the enthalpy and entropy of folding. Heat capacity measurements can also be used to indicate the state of the folded protein under varying conditions, and are simpler than alternatives such as FRET or NMR (Vajpai et al., 2013). Typically, an estimate for a heat capacity baseline is necessary to interpret the DSC results (Naganathan and Muñoz, 2014).

When folded into their native structure, proteins have a linear increase in their heat capacity with respect to temperature (Gomez et al., 1995; Privalov and Dragan, 2007; Naganathan et al., 2011b). This linear part of the heat capacity curve is the baseline, as shown in Figure 4.1. A reasonable estimate for the slope and the intercept of the baseline can be made by considering the molecular weight of the protein (Gomez et al., 1995; Muñoz and Sanchez-Ruiz, 2004) (the Freire baseline estimate). The Freire baseline has been used to determine whether a protein is folded or partially denatured at a given temperature (Naganathan et al., 2011b). It is worth noting that the Freire baseline underestimates the slope for DNA-binding proteins, which is usually explained by the increased flexibility of DNA-binding proteins (Privalov and Dragan, 2007).

The BICEP server, presented here, implements two methods to predict heat capacity baselines: the Freire baseline (Gomez et al., 1995) and a method based on the Hydrophobic Surface Area (HSA) using our recent modelling work (van Dijk et al., 2016b). This modelling work suggests that the slope of the heat capacity baseline should be directly correlated to the total amount of exposed hydrophobic surface area of a folded protein (van Dijk et al., 2016b); the model is based on current hydrophobic theory (Huang and Chandler, 2000; Widom, Bhimalapuram, and Koga, 2003b). In short, the hydrophobic effect is an effective force that has both an entropic and enthalpic component. This results in a temperature dependence of the hydrophobic force that may be approximated by a parabola, becoming weaker at high and low temperatures with a maximum at intermediate temperatures (Huang and Chandler, 2000; Widom, Bhimalapuram, and Koga, 2003b; van Dijk, Hoogeveen, and Abeln, 2015). Using a slightly modified derivation from the one published in Ref. (van Dijk et al., 2016b) to determine the enthalpic component of the free energy (see text S1 for derivation), we can show that the quadratic approximation leads to a linear relation between the heat capacity and the temperature.

We found that the hydrophobic surface area does indeed correlate with the slope of the heat capacity with respect to the temperature for a wide range of proteins, including the DNA-binding proteins (van Dijk et al., 2016b). For these latter proteins the high slope in the baseline may be explained by their large

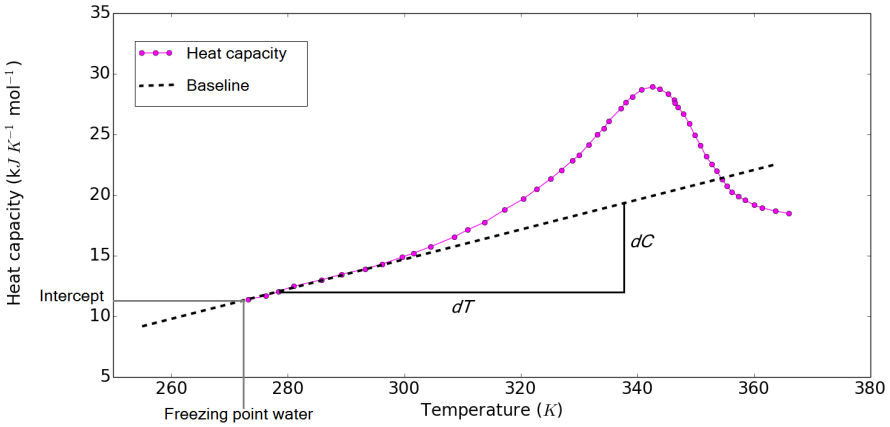

FIGURE 4.1: **Heat capacity curve obtained from a Differential Scanning Calorimetry (DSC) experiment.** In this experiment the heat capacity was measured for a range of different temperatures. For low temperatures the protein is in a native state. In this region, the temperature-heat capacity relation is approximately linear (the increase in the baseline of the heat capacity is illustrated by a dotted line). At intermediate temperatures, the heat capacity shows a sharp peak (reproduced from (Privalov et al., 1999).)

hydrophobic surface area. However, the published model is coarse-grained, and therefore not suitable for quantitative prediction of the heat capacity baseline for real proteins; instead it suggests how the constants may be fitted with help of experimental measurements. Here we set out to make this work practically available for the interpretation of DSC experiments.

The HSA method accurately predicts the slope of the heat capacity baseline with respect to the temperature for a wide range of proteins. On the other hand, the intercept is better modelled by the molecular weight. We make this method available through a web server that provides the user with the mathematical form of the baseline, along with the calculated HSA and molecular weight of the protein. We show in two separate use cases how the output may be used to interpret experimental baselines and compare the findings to earlier interpretations (Naganathan et al., 2011b; Naganathan and Muñoz, 2014).

Methods

Baseline estimation

In order to make the baseline estimate we run DSSP (Kabsch and Sander, 1983) over a set of PDB files corresponding to a set of protein structures for

which experimental heat capacity measurements are publicly available (see S1 Table).

The following equation is used to calculate the total hydrophobic surface area:

$$\begin{aligned} HSA &= \sum_{h_i \in N} g(h_i) \\ N &= \{A, F, C, L, I, W, V, M, Y\} \end{aligned} \tag{4.1}$$

The HSA is calculated by summing over the accessible surface area of all hydrophobic amino acids in a chain, where the set of hydrophobic amino acids (N) is defined as in Ref (Abeln and Frenkel, 2011). The function g returns the accessible surface area for a specific amino acid as calculated by DSSP. In case the protein is bound to DNA in the crystal structure, the DNA molecule is ignored in the calculation. Since the calorimetry measurements are also performed for free proteins, they allows for a proper comparison between experiment and prediction.

The HSA calculation is performed using eqn. (5.1). The baseline can be described by a simple linear equation:

$$C = a \cdot T + b \tag{4.2}$$

where C is the heat capacity in J mol$^{-1}K^{-1}$, T the temperature in degrees Celsius, a the slope of the baseline ($\frac{dC}{dT}$) and b the intercept (Figure 4.1). The intercept of the proteins that were used during the fitting procedure of the above formula are measured at 0 °C (Privalov and Dragan, 2007).

For the Freire baseline, eqn. (4.2) can be written as:

$$C = k_{slope,freire} \cdot M_{\text{weight}} T + k_{intercept,freire} \cdot M_{\text{weight}} \tag{4.3}$$

here, M_{weight} is the molecular weight in u. In ref. (Gomez et al., 1995), the constants $k_{slope,freire}$ and $k_{intercept,freire}$ were fitted to 0.0067 and 1.323 respectively. Since this work was published, more data on DNA-binding and globular proteins has become available. We provide the Freire baseline with the original constants for consistency with previous work. It is therefore not valid for DNA-binding proteins.

The HSA baseline can be written as:

$$C = k_{slope,hsa} \cdot HSA \cdot T + k_{\text{intercept},hsa} \cdot M_{\text{weight}} \tag{4.4}$$

here, the HSA is calculated as in eqn. (5.1). The constants $k_{slope,hsa}$ and $k_{intercept,hsa}$ are fitted to 0.090 and 1.31, respectively.

Web server

BICEP needs a Protein Data Bank (PDB) file as input. DSSP is used to calculate the surface area per residue. The user can supply a PDB identifier or their own PDB file. Optionally, the user can specify chains to analyse. Note that the HSA baseline prediction has only been validated on single chain structures. Optionally, experimental data can be visualized in the output plot by uploading a heat capacity profile.

The Freire baseline estimate that is solely based on molecular weight can also be plotted by selecting the check box. The user can also specify a point in the graph that should be intersected by the calculated equation. The point should be specified by giving the temperature (in K) and the corresponding heat capacity (in kJ mol$^{-1}K^{-1}$).

MatplotLib v 1.3.1 (Hunter, 2007) creates the figure that shows the baseline increase in heat capacity. PyMOL (DeLano, 2002) is used to visualise the protein with the hydrophobic residues coloured red and the remaining residues blue. The BICEP method has been implemented in Python and its source code is available at: `https://bitbucket.org/Robbinbwm/bicep`. We have made all functions available through a web server: `http://www.ibi.vu.nl/programs/bicepwww/`.

Results & Discussion

The baseline for the native structure of a protein can be described by two parameters: the slope of the baseline and the intercept of the baseline, noted by a and b respectively in eqn. (4.2). In this work, we investigate if these parameters are best fitted using the HSA or the molecular weight of a protein. To find the best predictor for the slope of the heat capacity, the correlation between the HSA and the slope of the heat capacity was compared to the correlation of the molecular weight and the slope of the heat capacity for a set of reference proteins (Privalov and Dragan, 2007) (see Figure 4.2 (a) and (b)).

Comparing Figures 4.2 (a) and (b), it can be observed that the correlation between the HSA and the slope of the heat capacity is higher than the correlation between the molecular weight (Freire baseline) and the slope, as reported in (van Dijk et al., 2016c). Moreover, when using the HSA there is no need to separate DNA-binding proteins and "globular" proteins in order to get a good correlation. This can also be observed by repeating the fit using only globular proteins (see Figure 4.2S). Note that the black line shown in Figure 4.2 (b) is drawn as given by (Gomez et al., 1995) and does not include DNA-binding proteins (coloured white), whereas all correlation statistics shown above the plots are reported for the entire dataset. Thus, the hydrophobic surface area model appears to be sufficient to make a reasonable estimate of the slope of the baseline for globular and DNA-binding proteins (van Dijk et al., 2016b): the high slope of the DNA-binding proteins may simply be explained by a large

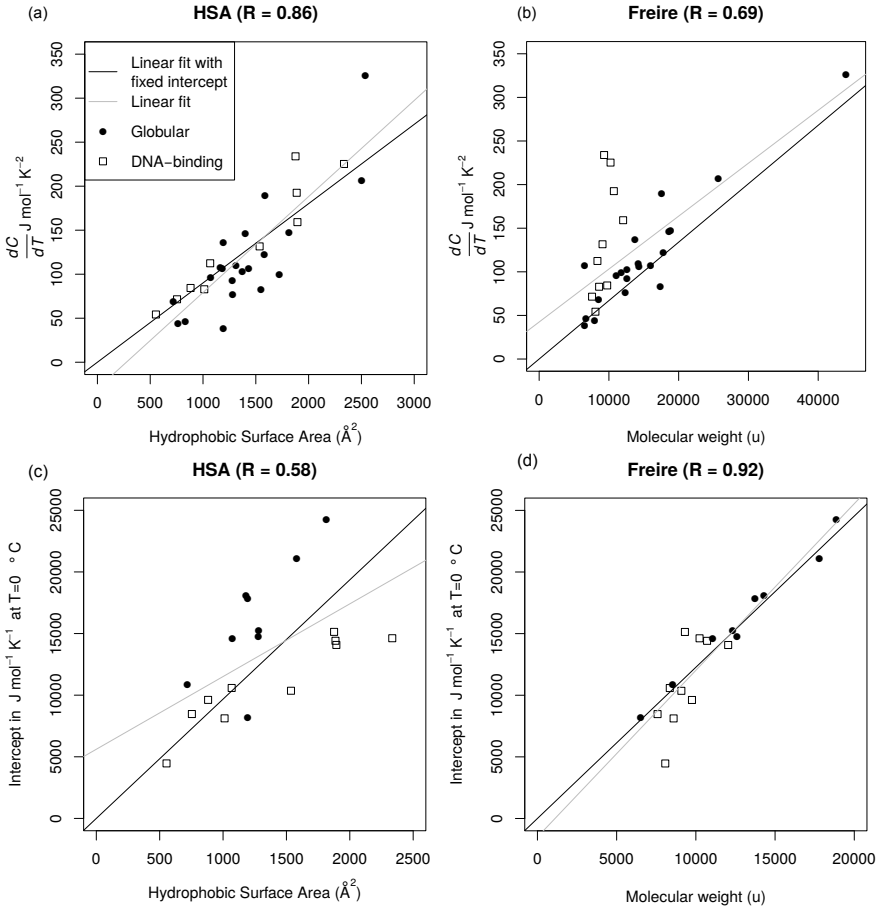

FIGURE 4.2: **Intercept and slope estimates for the heat capacity baseline**. Fit between the hydrophobic surface area (a), the molecular weight (b) and the slope of the heat capacity; fit between the hydrophobic surface area (c), the molecular weight (d) and the intercept of the heat capacity at 273K. Note that the black, solid line in (b) indicates the line originally fitted by Gomez et al. (Gomez et al., 1995) and does not take into account data determined after 1995. In particular, none of the DNA-binding proteins are included in this fit, nor some of the more recently determined globular proteins. Figure 4.2S shows the fit that includes all proteins. The R-value shown in (b) is based on the fit with all proteins. For the original dataset containing only nine globular proteins, the correlation was 0.90. The hydrophobic surface area shows a good fit for the slope on the full range of protein structures (both DNA-binding and globular proteins). The molecular weight is a better predictor for the heat capacity at 273K.

amount of hydrophobic surface area. This does not, however, preclude a possible correlation between the flexibility and the amount of exposed (hydrophobic) surface area.

A similar comparison was made for the intersection with the y-axis, which corresponds to the heat capacity at 0 . In this case, the molecular weight of the protein shows the highest correlation with the intercept (see Figure 4.2 (c) and (d)). The two parameters for the BICEP-baseline, a and b in eqn. 2, are determined using the best fit for both of them. This means that the HSA (Figure 4.2(a)) is used for the slope, a, and that the molecular weight (Figure 4.2(d)) is used for the intercept, b.

Above, we used fits that are fixed through the origin in Figure 4.2 (b) and (d). A protein with a molecular weight of zero should behave the same as the reference solution, and therefore have a heat capacity of 0 for every temperature. This corresponds to parameters of $a = 0$ and $b = 0$. This assumption is tested by fitting the data without this constraint (grey lines in Figure 4.2 (a) and (c)). As expected, the fits do not significantly change.

Based on our previous results in Ref. (van Dijk et al., 2016b), we would expect the slope of the heat capacity to be equal to zero in the temperature regime of a stable, folded protein if there is no exposed hydrophobic surface area. To test whether this prediction holds we also made a fit that was not fixed through the origin, see grey line in Figure 4.2 (a). As the fit does not significantly change, this result is consistent with our model. For the fit between the HSA and the intercept of the heat capacity, this argument cannot be made, since a protein with an entirely polar surface area could still have a heat capacity higher than that of the reference solution. This prediction is in agreement with the results, since the only combination with an intercept is significantly higher than zero is the fit of the HSA against the heat capacity at $0C$ (see Figure 4.2 (c)).

Heat capacity measurements may be affected by phenomena such as structural variations around the native state, aggregation or oligomerisation and molten globule ensembles, which in turn may affect our results. In the temperature regime below the unfolding transition, structural fluctuations in the native state may affect heat capacity measurements; such fluctuations generally result in a higher (slope of the) heat capacity. Ideally, only "well-behaved" proteins would be included in our analysis. To test whether the fit is significantly altered by the exclusion of "unideal" proteins in the low temperature range, we have excluded four proteins known to have transitional states between the folded and the unfolded state. These four proteins are marked in Figure 4.1S. A comparison of the fit without the four proteins shows that the difference between the fits is minor and unlikely to affect the conclusions of anyone using the BICEP-server for analysis.

Another problem is the possible oligomerisation or aggregation of proteins that may affect heat capacity measurements. One way to detect and correct for this is by determination of the *absolute* heat capacity, which uses measurements at different concentrations of protein (Kholodenko and Freire, 1999).

This method requires knowledge of some additional parameters, amongst which is the molar volume of a protein. Our analysis, based on data in Refs. (Gomez et al., 1995; Privalov and Dragan, 2007; Naganathan et al., 2011b), are based on measurements of the *apparent* heat capacity. While the *apparent* heat capacity is less accurate, there appears to be no systematic bias in the differences between the apparent heat capacity and the absolute heat capacity (Kholodenko and Freire, 1999). Moreover, in Ref. (Kholodenko and Freire, 1999) only a very small difference was found between the apparent heat capacity and the absolute heat capacity, far smaller than the error estimated by our fitting procedure.

At temperatures above the unfolding transition several effects, such as aggregation or molten globule formation, are very likely to affect the structural ensemble of the unfolded state and thereby influence (the slope of) the heat capacity. The baseline for the unfolded state can be of interest alongside the slope of the native baseline. In addition, recent theory (van Dijk et al., 2016b) suggests a linear relation between the temperature and the heat capacity for the unfolded state, that correlates to the hydrophobic surface area of the protein. Unfortunately, no suitable dataset is available to confirm this. Moreover, it would be difficult to draw any conclusions due to the diversity of structural ensembles in this temperature range. Thus, we cannot validate predictions on the slope of the heat capacity in the unfolded regime, and therefore do not include such predictions in our web server.

PDD - fast-folding protein

The fast-folding protein PDD is frequently used in folding studies and models because of its fast folding property and small size (Naganathan and Muñoz, 2014). Two separate experimental techniques to determine secondary structure suggest that the protein gradually unfolds from 280K to 350K (Naganathan and Muñoz, 2014; Naganathan et al., 2010).

The heat capacity obtained through DSC-experiments (purple dots Figure 4.3) was higher and with a steeper slope than would be expected from the Freire baseline (red line Figure 4.3). The Freire baseline yields the following constants in eqn. (4.2): $a = 30.69 \pm 13$ J mol^{-1} K^{-2} and $b = 6060 \pm 540$ J mol^{-1} K^{-1}.

The HSA-based method finds alternate values and error margins for the constants in the same equation for the PDD protein: $a = 107.68 \pm 18.9$ J mol^{-1} K^{-2} and $b = 6000 \pm 1353$ J mol^{-1} K^{-1}.

In contrast to the Freire baseline, the HSA-based method does seem to predict the slope and intercept of the heat capacity accurately (Figure 4.3). The denaturation peak of the protein is inside the error bars for the HSA-based method. This supports the hypothesis (Naganathan and Muñoz, 2014; Naganathan et al., 2010) that the protein structure unfolds gradually between 280K and 350K.

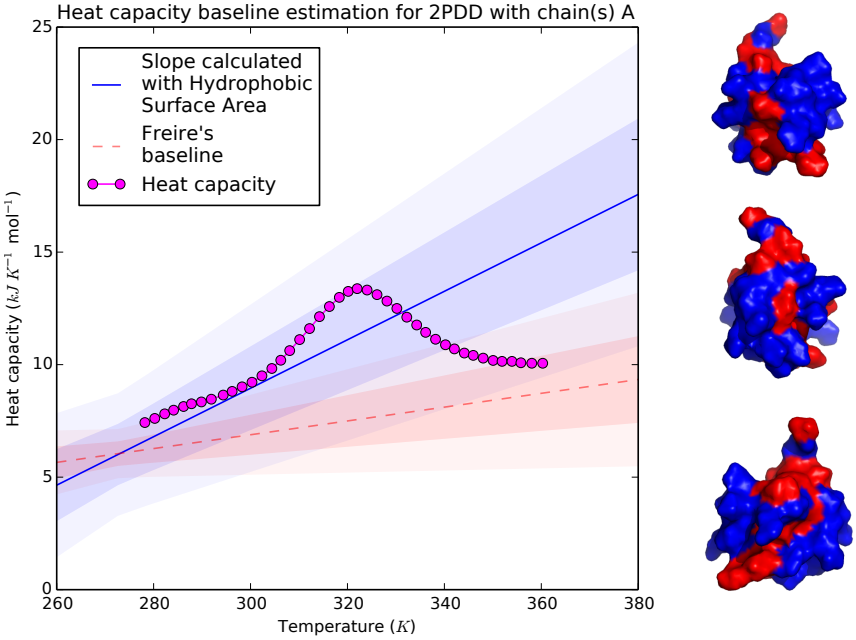

FIGURE 4.3: **Output BICEP server for PDD.** The left panel shows the baseline estimation with the HSA-based method and Freire baseline (Kalia et al., 1993). One and two standard deviations are indicated by shaded areas of different opacity. The structure of the PDD (PDB identifier: 2PDD) protein. From NMR-measurements can be concluded that the terminal regions are largely unstructured (Kalia et al., 1993). The Hydrophobic Surface Area visualized for the PDD protein with the PDB identifier 2PDD, viewed from three different angles. The hydrophobic amino acids are coloured in red and the remainder in blue. The set of hydrophobic amino acids is defined in eqn. (5.1). The protein is shown in three configurations: rotated 0° (top), 120° (middle) and 240° (bottom).

PDD has a relatively large HSA in the PDB-structure (Figure 4.3), which explains why the HSA-based method predicts a large slope for the baseline. Investigation of the protein structure indeed shows that the terminal regions are largely unstructured (Kalia et al., 1993), which makes it more likely that a large part of the hydrophobic surface is exposed to the solvent; this explains why the slope of the BICEP-baseline is larger than the slope of the Freire-baseline.

SRY - DNA-binding protein

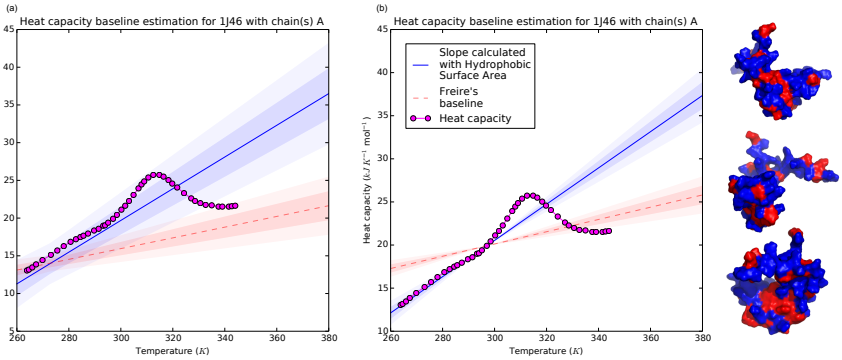

FIGURE 4.4: **Output BICEP server SRY.** Panel (a) shows the HSA slope based method and Freire baseline (Murphy et al., 2001). Panel (b) shows the Freire baseline with a fixed heat capacity at 308 K. The first and second standard deviation are plotted with a different opacity. On the right, the Hydrophobic Surface Area is visualized for the SRY protein with the PDB identifier 1J46, viewed from three different angles. The hydrophobic amino acids are coloured in red and the remainder in blue. The set of hydrophobic amino acids is defined in eqn. (5.1). The protein is shown in three configurations: rotated 0° (top), 120° (middle) and 240° (bottom).

The human male sex-determining factor SRY (PDB identifier: 1J46) is a DNA-binding protein. The slope of the heat capacity for DNA-binding proteins is higher than one would expect based on the Freire baseline (Privalov and Dragan, 2007; Naganathan et al., 2011b), see Figure 4.4 (a). The experimental data for SRY was taken from ref. (Dragan et al., 2004).

The Freire baseline yields the following constants in eqn. (4.2) for the SRY-protein: $a = 71.12 \pm 13$ J mol^{-1} K^{-2} and $b = 14044 \pm 540$ J mol^{-1} K^{-1}. The HSA-based method yields different constants: $a = 210.22 \pm 18.9$ J mol^{-1} K^{-2} and $b = 13025 \pm 1353$ J mol^{-1} K^{-1}. Hence the slope of the heat capacity is higher than the estimate given by the Freire baseline, as is the case for most DNA-binding proteins. The derived fit, which is valid for both DNA-binding and globular proteins, yields a better baseline prediction for this protein (van Dijk et al., 2016b), see blue line in Figure 4.4 (a); this may again be explained by the rather large amount of hydrophobic surface area in the DNA-binding pocket of the protein.

Figure 4.4 (a) also shows that, similar to the PDD use case, the unfolding peak of the protein lies inside the error estimates for the HSA-based method. Here both baselines are shown with two standard deviations (shaded) and experimental data is also provided (circles). It has been observed experimentally

that the protein gradually unfolds (Dragan et al., 2004) which may lead to a relatively low heat capacity peak.

To get a more realistic estimate of the error bound for the heat capacity, we fix the heat capacity at T=308 K, to reduce the uncertainty due to the estimate of the intercept (see Figure 4.4 (b)). Here we assume that the heat capacity and the structure at 308 K are correctly measured. Comparing the baseline estimate with the experimental data suggests that the protein behaves as expected from the total amount of hydrophobic surface area calculated from the structure determined at 308K in the linear regime. The (heat) denaturation peak now falls outside the margin of error, signifying that there is an unfolding event at approximately 315K.

Conclusion

Here we aim to facilitate the interpretation of DSC-curves using current hydrophobic theory. For the Freire baseline, the molecular weight of a protein is used to calculate the slope and the intercept of the baseline. It was shown by our group that the slope of the heat capacity baseline can be more accurately predicted by an estimate based on the hydrophobic surface area of the folded structure (van Dijk et al., 2016c), whereas the intercept is most accurately predicted by an estimate based on the molecular weight. The BICEP web server implements the Freire baseline, as well as the baseline developed in our group, which is valid for a wider range of proteins. This web server provides an easily accessible way to analyse and visualise the results of these experiments and is available at `www.ibi.vu.nl/programs/bicepwww`.

Acknowledgments

SA has been supported by a Veni grant on the project 'Understanding toxic protein oligomers through ensemble characteristics' from Netherlands Organisation for Scientic Research (NWO).

4.1 Supporting Information - BICEP: Heat Capacity Baseline Prediction

S1 Text

Derivation of the relation between heat capacity and hydrophobic surface area. In this section we present a derivation of the relation between the heat capacity and the hydrophobic surface area. The heat capacity can be calculated as follows:

$$C(T) = \frac{d\langle E\rangle}{dT} \tag{4.5}$$

As in ref. (van Dijk et al., 2016b), we write the free energy F as the sum of the potential energy from pairwise amino acid interactions (E_{int}) and hydrophobic interactions ($F_{\text{hydrophobic}}$):

$$F = E_{\text{int}} + F_{\text{hydrophobic}} \tag{4.6}$$

Here, the $F_{\text{hydrophobic}}$ is approximated with a quadratic potential that models the temperature dependency of hydrophobic interactions. Here, T is the temperature, T_0 is the temperature where the free energy for hydrophobic interactions is maximal (E_0) and the temperature dependency is multiplied by a constant, α. In ref. (van Dijk et al., 2016b) the derivation is performed for a lattice model, and uses the same temperature dependence for all amino acids. Here we adapt this method to take into account the differing amounts of exposed hydrophobic surface. This is done by replacing the number of hydrophobic amino acids at the surface with the HSA, indicated by σ here. This results in the following relation:

$$F = E_{\text{int}} + \sigma(E_0 - \alpha(T - T_0)^2)) \tag{4.7}$$

The expected total energy of the system $\langle E\rangle$ can be calculated by introducing β or $\frac{1}{T}$ in the equation:

$$\beta F = \beta E_{\text{int}} + \beta\sigma(E_0 - \alpha(T - T_0)^2) \tag{4.8}$$

By taking the derivative $\frac{d\beta F}{d\beta}$ the internal energy ($\langle E\rangle$) is calculated:

$$\frac{d\beta F}{d\beta} = \langle E\rangle = E_{\text{int}} + \sigma E_0 - \sigma\alpha(-\frac{1}{\beta} + T_0^2) \tag{4.9}$$

This equation can be simplified to:

$$\langle E\rangle = E_{\text{int}} + \sigma E_0 - \sigma\alpha(T_0^2 - T^2) \tag{4.10}$$

By taking the derivative of this equation we can calculate the heat capacity:

$$\frac{d\langle E\rangle}{dT} = C = 2\sigma\alpha T \tag{4.11}$$

The equation shows that the HSA (σ) can be used to estimate the heat capacity (C), and therefore the native baseline. The equation also shows that the heat capacity baseline can be estimated using a linear equation with respect to temperature.

This derivation does not take into account the freezing of water, so in practice we use the heat capacity at 0 in combination with a linear fit of the heat capacity, to find a prediction of the heat capacity.

S1 Table

Dataset of protein structures Experimental data gathered from ref. (Privalov and Dragan, 2007) for the linear increase of the heat capacity baseline.

PDB	Slope $\frac{dC}{dT}$ $(J{\cdot}\mathrm{mol}^{-1} \cdot K^{-2})$	Intercept at 0°C (kJ)	Length	Hydrophobic Surface Area (Å^2)	Molecular Weight (u)	Type
5PTI	37.7	141	58	1194	6525	globular
1RNB	76.3	140	109	1281	12328	globular
1MBO	122.4	138	153	1581	17801	globular
4II8	105.78	140	129	1182	14307	globular
5CYT	92.7	143	103	1279	12600	globular
1UBQ	68.4	143	76	719	8539	globular
3LZM	147.6	148	164	1814	18860	globular
8RNT	95.68	140	104	1074	11068	globular
7RSA	136.4	144	124	1193	13725	globular
3HDD	71.5	154	55	754	7583	DNA
1APL	84.37	163	59	882	9755	DNA
9ANT	82.88	145	56	1012	8595	DNA
1HMA	112.42	145	73	1068	8354	DNA
2ZTA	54.25	144	31	554	8070	DNA
1I11	131.6	148	70	1536	9085	DNA
1TF3	159.16	153	92	1894	12040	DNA
1CG7	192.51	155	93	1887	10708	DNA
1J46	225.25	172	85	2335	10234	DNA
2LEF	233.92	176	86	1876	9310	DNA
1RIL	145.5693	NA	146	1400	18621	globular
1F21	189.4832	NA	152	1586	17552	globular
4HU7	99.4244	NA	107	1724	11760	globular
1DBY	102.396	NA	112	1372	12600	globular
1HFX	109.4823	NA	123	1313	14200	globular
1SYD	106.6648	NA	136	1432	16000	globular
8I1B	82.48375	NA	151	1550	17391	globular
1YO7	107.364	NA	120	1163	6510	globular
1E6H	46.08855	NA	61	834	6700	globular
1HDN	43.792	NA	85	764	7960	globular
1OVA	325.7766	NA	366	2537	44000	globular
1EX3	206.24	NA	245	2502	25686	globular

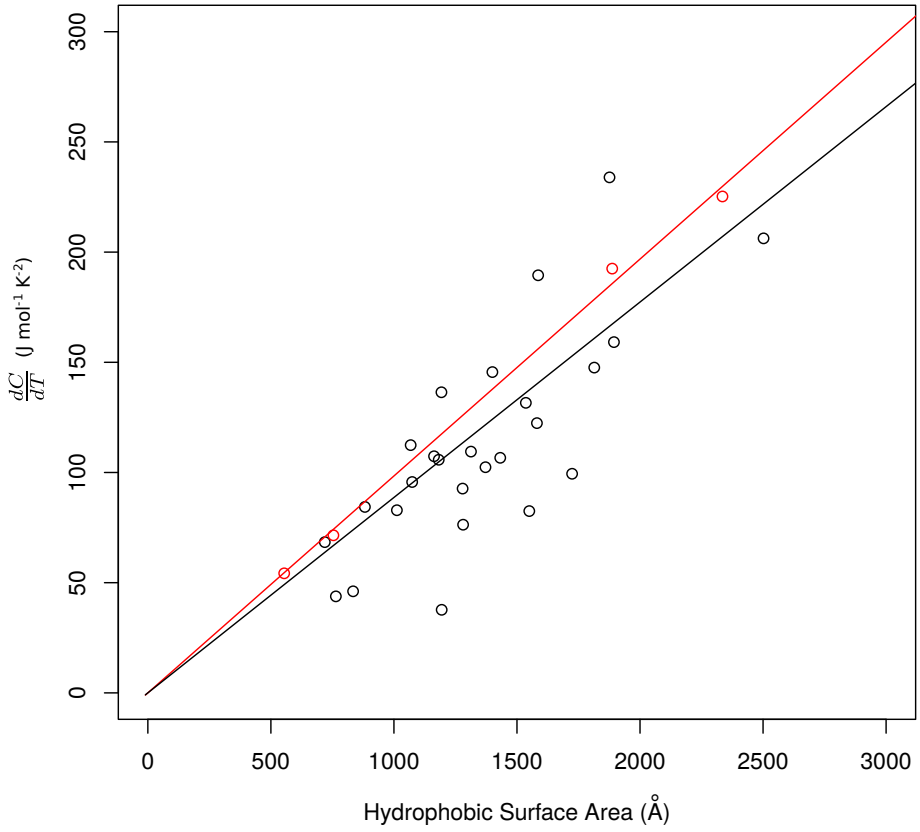

FIGURE 4.1S: **Effect of natively disordered proteins on the fits.** Fit between the hydrophobic surface area and the slope of the heat capacity. The red dots show proteins that do not have a single, well-defined native state, but instead unfold in multiple stages. The black dots show "normal" proteins. The solid line shows the same fit as in the main text, which includes all proteins in the dataset, while the dashed line shows the fit where proteins with multiple unfolding transitions are excluded. The difference between the two fits is small.

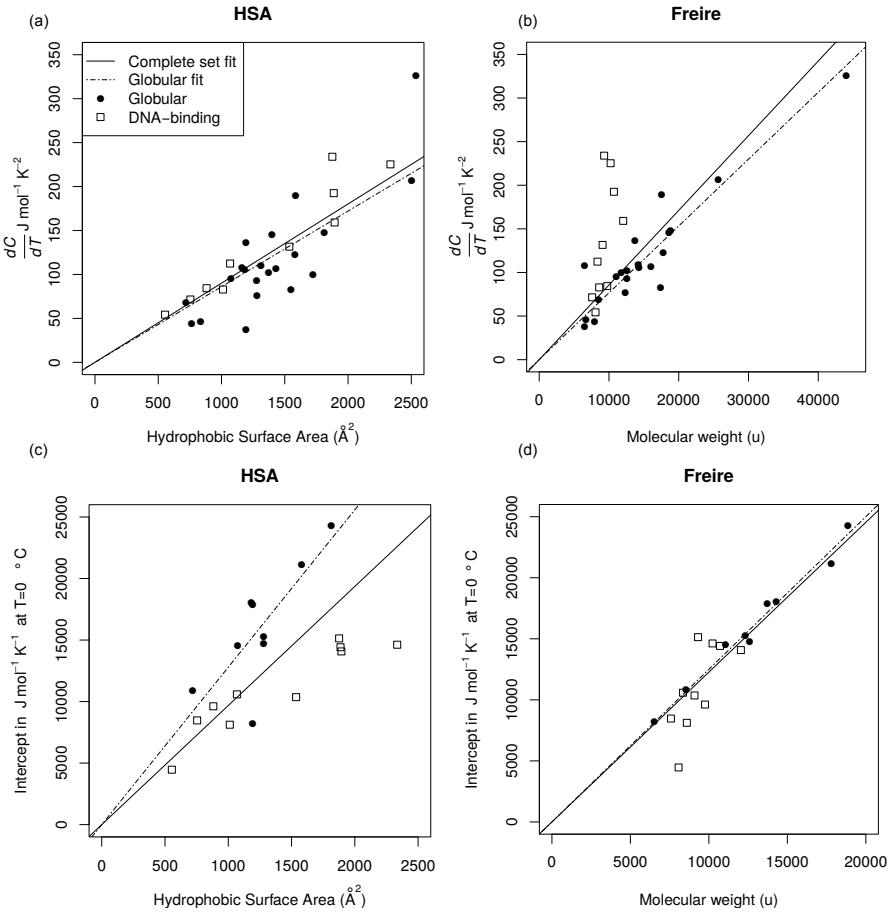

FIGURE 4.2S: **Effect of proteins with and without well-defined native state on baseline estimate.** Fit between the hydrophobic surface area (a), the molecular weight (b) and the slope of the heat capacity; fit between the hydrophobic surface area (c), the molecular weight (d) and the intercept of the heat capacity at 273K. The hydrophobic surface area shows a good fit for the slope on the full range of protein structures (both DNA-binding and globular proteins). The molecular weight is a better predictor for the heat capacity at 273K. For all graphs, the solid black line indicates the fit made with the entire data set, while the dashed line indicates the fit made with only the globular proteins. The fits we used in our final method, shown in (a) and (c), do not depend on the type of protein investigated, in contrast to the other fits shown in (b) and (d).

Chapter 5

Predicting the hydrophobic surface area of native protein structures from sequence

Based on the manuscript:

R Bouwmeester, E van Dijk, J Heringa, and S Abeln (2016). "Predicting the hydrophobic surface area of native protein structures from sequence". In: *In preparation*

Abstract

Proteins tend to bury hydrophobic residues inside their core during the folding process. Nevertheless, for many proteins a small fraction of the hydrophobic residues remains on the surface. These hydrophobic patches often play an important functional role, for example in protein-protein interactions, ligand binding and interactions with the membrane. There are several methods that, from the protein sequence, predict whether a particular residue will be exposed to the surface. Hydrophobic residues will typically be predicted as buried by such methods, following the general trend. Hence, predicting the total solvent accessible hydrophobic surface from the amino acid sequence remains a challenging problem. This problem is highly relevant, for example to estimate aggregation behaviour for a specific protein and or estimate the interaction strengths with a hydrophobic surface as used in many experimental setups. Here, we present a method that predicts the total amount of hydrophobic surface area. We show that this approach estimates the total hydrophobic surface area more accurately than predictions based on single residue surface area predictions, even though the model is extremely simple and only uses the protein sequence length, the number of hydrophobic and hydrophilic amino acids as features.

The simplicity of the presented model - which does not take evolutionary conservation into account - indicates that basic physical constraints determine the total amount of hydrophobic surface. Moreover, taking these simple physics based features into account may help to improve the accuracy of other surface based prediction methods; we show for example that our model can be used to improve the accuracy of accessible surface area prediction per residue.

Introduction

Hydrophobic amino acids are generally buried in the protein core upon folding. This is known as the hydrophobic collapse and is the driving force of the folding process (Wiggins, 1997). However, some hydrophobic residues are present on the surface of the protein (Figures 5.1 and 5.1S) (Kato and Nakai, 1980; Eisenhaber, 1996; van Oss, 1995; Scarsi, Majeux, and Caflisch, 1999). The hydrophobic residues on the surface of proteins often fulfill important functions in protein-protein interactions, DNA-binding, and ligand binding. A simple way to quantify the amount of surface exposed hydrophobic residues, is by counting the total hydrophobic surface area (HSA) on the surface of the protein, see Figure 5.1.

The total hydrophobic area that is exposed to the surface of a protein has important consequences for the solubility of the protein, both in cellular and experimental context. Non-specific aggregation of proteins (Beyreuther et al., 1991; Ross and Poirier, 2004; Nieba et al., 1997), protein-protein interactions (Chothia and Janin, 1975; Young, Jernigan, and Covell, 1994) and misfolding (Dobson, 2004; Branden and JohnTooze, 1999) are directly related to the amount of hydrophobic surface area.

Moreover, many experimental techniques take advantage of the fact that some proteins have a more hydrophobic surface area than others, such as gel electrophoresis (Wilkins et al., 1998) and high performance liquid chromatography (HPLC) (Kaliszan, 1990). HPLC is used to separate samples based on the retention time of the peptides or proteins under investigation. Knowledge of the retention time before the start of the experiment can be used to improve metabolite detection in HPLC methods (Boswell et al., 2011). Typically, the hydrophobicity of the peptide or protein correlates strongly with the retention time, thus knowledge of the hydrophobicity of a peptide or protein can improve the sensitivity of metabolite detection.

Furthermore, the total hydrophobic surface area may also directly influence the measured experimental outcome. A good example is differential scanning calorimetry (DSC), for which the heat capacity temperature relation for the folded protein is directly related to the total hydrophobic surface area (Gómez et al., 1995; van Dijk et al., 2016c). In fact, the work in (van Dijk et al., 2016c) gave the motivation to obtain the aggregate hydrophobic surface area from a protein sequence.

In contrast to the above methods, where the hydrophobic surface area is a parameter in the experimental setup, for other experimental techniques, proteins with a large hydrophobic surface can hinder the experiment. For example, a large hydrophobic surface area can lead to gel formation in protein crystallization experiments (Durbin and Feher, 1996; Wright and Dyson, 1999; Tusnády, Dosztányi, and Simon, 2005).

To predict whether a particular residue will be exposed to the surface, given the amino acid sequence alone is still a challenging problem; we will refer to this problem as per-residue surface area prediction. Several methods exist:

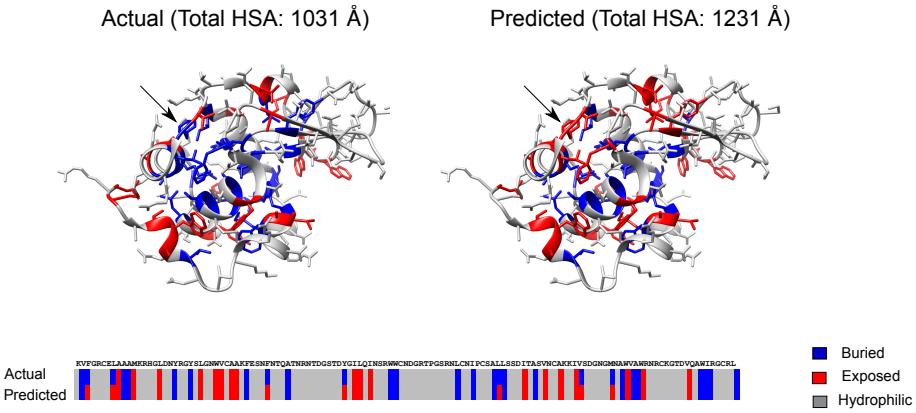

FIGURE 5.1: Predicting surface exposed hydrophobic residues.The left panel shows structure-assigned accessibility annotations for hydrophobic residues for lysozyme (PDB-id: 5EBH), if an amino acids has less than 10% exposed solvent area, it is classified as buried (blue), otherwise it is shown as exposed (red). The right panel shows predictions for the hydrophobic residues as provided by NetSurfP, with residues predicted to be on the surface in red, and those predicted to be buried in blue. Comparing the predicted and structure-assigned accessibility annotations, it is clear that the majority of hydrophobic residues are predicted to be buried. In contrast, two of the hydrophobic amino acids that are predicted to be exposed, are in fact buried (indicated by arrow). Though just a single example, it appears to be a difficult problem to predict which hydrophobic amino acids will reside on the surface of the protein upon folding.

SANN (Joo, Lee, and Lee, 2012), SARPRED (Garg, Kaur, and Raghava, 2005), SPINEX (Faraggi et al., 2012) and NetSurfP (Peter, 2010). All these prediction tools use a windowed approach and a Position Specific Scoring Matrices (PSSM) as feature vector. SPINEX adds physical properties, like a hydrophobicity index. Secondary structure predictions are included in a similar approach for SPINEX. The tools SPINEX, NetSurfP and SARPRED all use Neural Networks (NN); NetSurfP and SARPRED use an ensemble of layered NNs. SANN is the only tool that uses the K-Nearest Neighbour algorithm to predict the hydrophobic surface area.

The above methods can predict the surface exposure of amino acids with a reasonable accuracy. Generally, these methods will predict the burial of hydrophobic amino acids and exposure of hydrophilic acids, see Figure 5.1. This means that using these predictions to calculate the total accessible hydrophobic surface area will not necessarily provide robust results, even though exactly this information would be highly relevant for predicting the propensity of proteins to interact with other biomolecules or aggregate.

The importance of the HSA in experimental methods and biological phenomena raises the question: can we predict the HSA, without making explicit predictions for each individual residue? We show here that a simple model, based on physical principles, provides a better estimate for the HSA than prediction based on individual residues. The model implicitily relies on the observation that globular proteins are approximately spherical (Chothia, 1976b; Abeln, 2007). This allows the model to predict the surface area, without exactly knowing what individual amino acids are exposed on the surface.

The regression model proposed here uses the total surface area in combination with the amount of hydrophobic residues to predict the HSA. Note that this model predicts the total hydrophobic surface area instead of determining which specific residues are present on the surface. Nevertheless, we investigate whether the latter type of predictions may be improved using similar global sequence features.

Methods

Structure assigned hydrophobic surface area

In order to validate the predictions, we first need a measure for the total hydrophobic surface area. Here we take a simple definition, similar to a definition that has previously shown to correlate well with physical observables (van Dijk et al., 2016c).

We define the total HSA as the sum over the accessible surface area (ASA) of all hydrophobic residues (Hydr) for a proteins:

$$HSA = \sum_{h_i \in \mathrm{Hydr}} g(h_i) \tag{5.1}$$

here $g(h_i)$ is the accessible surface area calculated by DSSP (Kabsch and Sander, 1983) for a hydrophobic residue, h_i. DSSP derives its accessibility values from three-dimensional, structural information as deposited in the Protein Data Bank Berman et al., 2000b. The hydrophobic residues are in the work defined as:

$$\text{Hydr} = \{A, F, L, I, W, V, M, Y\} \tag{5.2}$$

i.e. the amino acids Ala, Val, Ile, Leu, Met, Phe, Tyr, Trp were considered hydrophobic. The polar (hydrophillic) residues are defined as:

$$\text{Pol} = \{K, R, H, D, T, E, N, Q\} \tag{5.3}$$

so the amino acids Lys, Arg, His, Asp, Ser, Thr, Glu, Asn, Gln are considered to be hydrophilic. Pro, Cys and Gly are considered special cases, and are therefore neither considered hydrophobic nor hydrophilic.

Predicting the HSA

In this work we discuss three approaches to predict the HSA from sequence: 1) A Global prediction approach. This to predict the hydrophobic surface area. These methods use features of the whole sequence, like length and the total hydrophobic residues to predict the HSA. 2) Purpose built per residue prediction methods. These methods make a per-residue prediction with the sole purpose of using it to predict the total HSA. 3) The adapted per residue methods. This approach uses an existing method to predict the surface area for each individual amino acid, and adds the predictions for the hydrophobic residues to obtain a prediction for the total HSA.

Models to predict the total HSA

The TFM model described in this work uses three features: sequence length, the number of hydrophilic and hydrophobic amino acids (eqn. 5.2) and was trained using a regression based approach. This will be called the Three Feature Model (TFM) throughout the text. The following machine learning algorithms were evaluated to obtain the parameters in the TFM model: Blackboost, Cubist, Neural networks, Extreme Learning Machine, spls, superpc, glmnet, Random Forest and Lars.

The TFM model was trained with Evtree and NodeHarvest. The latter algorithms are computationally expensive, therefore they were not used for the other two methods that have a higher number of features. The machine learning algorithms in the R-package CARET were used for training and parameter optimization (Kuhn et al., 2014). The average absolute deviation from the ASA according to DSSP was used to evaluate the windowed ASA prediction method.

Purpose-built per-residue surface prediction methods

We have investigated two methods to predict the total HSA of a protein structure using a purpose-built per-residue approach. The first method predicts the ASA per amino acid in the sequence using a sliding window. The HSA was then calculated using eqn. (5.1). The second model introduced a second layer to the first method by incorporating extra features retrieved from the entire protein sequence. The features in the second layer were the predicted HSA from the previously described model, sequence length and a vector of length 20 that contained amino acid frequencies. For a more detailed description of these methods, see SI.

Adapted per-residue surface prediction methods

Adapted per-residue surface prediction methods use existing prediction methods for single amino acids, and combine them to obtain the total HSA for a protein. We compared the total HSA-prediction from the TFM-model to predictions obtained from per-residue methods. We considered the methods SANN (Joo, Lee, and Lee, 2012), SPINEX (Faraggi et al., 2012), SARPRED (Garg, Kaur, and Raghava, 2005) and NetSurfP (Peter, 2010). The parameters of the ASA prediction tools were not optimized; default settings were used. These methods predict the accessible surface area per residue. In order to convert the predicted surface area per residue, to the total hydrophobic surface area we took the sum over the prediction for all hydrophobic residues. We have also investigated a more elaborate way of converting per-residue predictions to a surface area prediction.

Length-based reference model

Besides comparing to adapted per-residue surface area prediction methods, we also compared to a length-based reference model. This simple model consists of an estimate based on the length of the protein sequence. The idea of approximating proteins as a sphere to predict the ASA of the whole protein was first introduced in (Janin, 1979). The ratio between hydrophilic and hydrophobic residues on the surface has previously been observed in (Chothia, 1976b): for proteins with a high molecular weight the ratio of hydrophobic residues can be well approximated for globular proteins based on the length of the protein sequence alone.

The reference model uses the sequence length of a protein (L) multiplied with a constant (k_1) and to the power of a constant (k_2) to predict the HSA:

$$ASA = k_1 \cdot L^{k_2} \tag{5.4}$$

Note that in case of a perfect sphere, we would have:

$$\text{Surface area} = 4\pi((\frac{3V}{4\pi}))^{\frac{2}{3}} \tag{5.5}$$

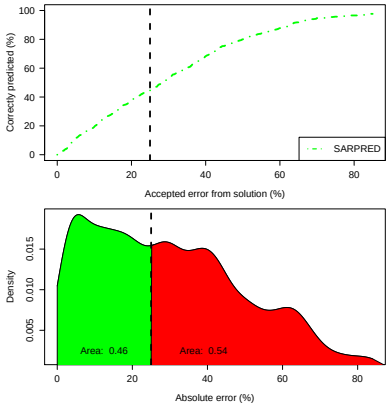

FIGURE 5.2: Explanation of evaluation metric for the HSA-prediction methods. The method SARPRED is used as an example. A range of different error thresholds are used to get the ratio of correctly predicted HSAs. A structure can have a correctly predicted HSA (absolute error $<$ threshold) or an incorrectly predicted HSA (absolute error$>$threshold). The percentage correctly predicted structures (The y-axis in panel (a)) was evaluated for a varying threshold (indicated by dashed line in both panels) corresponds to the percentage of correctly predicted structures within the accepted error from the solution, shown in by the green plane in panel (b). Ideally, a method would have a high density at a low absolute error in panel (b), and thus have a high proportion correctly predicted structures at low errors in panel (a)

Using the latter equation the total ASA can be approximated by assuming the sequence length (L) scales linearly with the volume (V). Since proteins are not perfect spheres and only a fraction of the surface is covered by hydrophobic groups, we instead generate the baseline model by fitting the constants k_1 and k_2 to the training set, minimising the sum of squares between the predicted and observed HSA. In this simple model the assumptions were made that the ratio of hydrophobic amino acids on the surface in relation to the length is constant.

Benchmarking

Dataset of protein structures used

The PDBselect database with 25 % sequence identity and nsigma = 3.0 (Griep and Hobohm, 2010) was used to train and validate the HSA prediction. The nsigma level is a statistic that is based on the sequence identity and the length of the sequence alignment (Abagyan and Batalov, 1997). It is used here to exclude short sequence alignments with high sequence identity that do not share a similar fold.

In addition to the above redundancy filter, 103 transmembrane proteins were removed (Lomize et al., 2006). Transmembrane proteins have a significantly higher hydrophobic surface than soluble proteins, which contradicts the key assumption in our model that hydrophobic amino acids tend to be buried. The presented model is optimised for predicting the HSA of soluble proteins. In the case of multiple chains in a protein structure the chains were split into structures with a single chain.

The filtered dataset consists out of 5,567 sequence chains with a total of 741,964 amino acids. The proteins were annotated with their publication date by parsing the "PDBx:data" field from their respective PDB files in XML format.

Dataset splitting

The filtered dataset was split in a training, validation and test set. The training set contains 60 % of the protein structures, the validation set contains 20 %, and the test set also contains 20 %. The split was performed randomly. The training set was used solely to train the TFM model, the length-based reference model, and the purpose-built per residue models. The validation set was used to select the best performing model. The test set was used to determine the performance of the best model.

The final model was trained on all available proteins, so the performance numbers using the training set represent a lower bound of the performance.

Correlation and mean error

The performance of the methods was evaluated using the Pearson correlation coefficient between predictions and the reference HSA determined with DSSP using eqn. (5.1). The measure gives an indication of the quality of the predictions. However, this measure has some disadvantages: the correlation is sensitive to outliers and does not give an indication of the error margin in the prediction. The mean error was used to indicate the average distance between the predicted HSA and the HSA calculated from the structure. The mean error is also sensitive to outliers, but is comparable to an error measure which is often used in per-residue prediction methods: the mean error per residue. To find the importance of the different features, we also calculated the performance of the models where various features were omitted from the model.

Relative error threshold curve

We use a relative accuracy threshold to define how well a prediction method performs, given a specific threshold, Figure 5.2 illustrates the relative error threshold curve. For each prediction, the relative error, δ_i, with respect to the true HSA, for a protein structure, i, is calculated as follows:

$$\delta_i = \frac{|\mathrm{HSA}_{\mathrm{predicted},i} - \mathrm{HSA}_{\mathrm{DSSP},i}|}{\mathrm{HSA}_{\mathrm{DSSP},i}} \tag{5.6}$$

The performance of the methods over the whole set of *structures* are evaluated by plotting the percentage correctly predicted instances (protein chains) versus a varying threshold, t. The threshold curve shows the percentage of correctly predicted HSA of proteins for a given relative error threshold:

$$F(t) = \frac{\#\{i|i \in \text{structures} \wedge \delta_i < t\}}{\#\{i|i \in \text{structures}\}} \tag{5.7}$$

Where F is the fraction correctly predicted residues within the threshold, t. In the curve, shown at the top panel of Figure 5.2 , F is plotted along the y-axis and t on the x-axis.

Results & Discussion

Here we set out to investigate whether it is possible to predict the total amount of (exposed) hydrophobic surface area from a protein sequence. We considered three general approaches: firstly, we constructed several methods that were trained to only predict the total hydrophobic surface area of a protein. Secondly, we built a per residue approach to predict the accessible surface area for each residue, and than combine the predictions for the hydrophobic residues to obtain the total HSA. Thirdly, we adapted several existing methods that predict the surface accessible area for specific residues to estimate the total hydrophobic surface area for a protein.

We first evaluate different methods for our first approach. The best performing method is the Three-Feature Model (TFM) based on sequence length and number of hydrophobic and hydrophilic amino acids. We also considered a length derived formula, based on eq. 5.4, to provide a simple lower bound for the prediction method (See Figure 5.3S). Note that the type of regression model used to train the TFM, had a small but significant influence on the method performance. The cubist algorithm gives the best performance (Figure 5.2S).

For the second approach, the purpose built per residue approach, we evaluated two methods. The first method predicts the ASA of specific amino acids using a simple windowed approach. The second method takes the HSA predictions from the simple windowed method and adds amino acid frequencies as features (Two-layer method), which are then combined using a machine learning method.

For the third approach, we adapted several methods to predict the accessible surface area: SANN (Joo, Lee, and Lee, 2012), SARPRED (Garg, Kaur, and Raghava, 2005), SPINEX (Faraggi et al., 2012) and NetSurfP (Peter, 2010). In order to convert the predicted surface area per residue, to the total hydrophobic surface area we simply took the sum over the prediction for all hydrophobic residues (see Methods for further details).

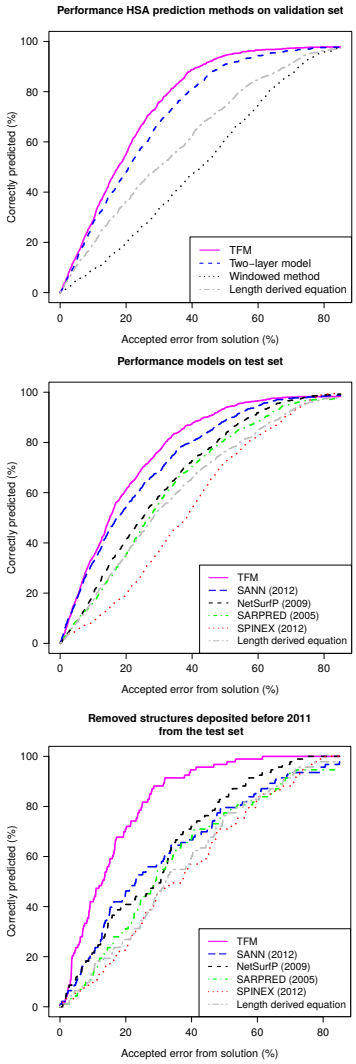

FIGURE 5.3: **Performance comparison methods. The ratio of correctly predicted HSAs ($F(t)$ in eqn. (5.7)) as a function of the error thresholds (t)** Panel (a) shows the performance of purpose-built methods developed to predict the total hydrophobic surface area from sequence information. Panel (b) shows a performance comparison of adapted methods and TFM. The performance comparison was run on the test set in (b). In (c) the evaluation was performed on 97 proteins from the test set, that were deposited after 2011 to the PDB. Removing these proteins allows us to get an idea of the amount overtraining of existing methods that were developed before 2011.

In order the compare the performance between the methods, we carefully split a filtered set of available protein structures into a training, test and validation set and compare the structure-assigned hydrophobic surface area to the predicted hydrophobic surface area. We use a threshold based approach to evaluate the accuracy of the different methods, see Methods for further details and an example. Figure 5.3 shows that, surprisingly, the simple three-feature based method (TFM) outperforms the window based methods.

For the published per-residue prediction methods, it is less easy to split the available protein structures in a training, validation and test set, because the methods have likely been trained on all available structures. Therefore, there is a risk that the methods have been overfitted to the data that was present in the Protein Database at the time of publication. To exclude this possibility, we used a set of structures that were released in the PDB (Berman et al., 2000b) after 2011 for validation. Since all the per-residue methods investigated here were released before this date, these structures provide a good validation set. Surprisingly, the very simple three feature method outperforms all existing methods for almost all reasonable error thresholds, as shown in Figure 5.3 (c).

Table 5.1 shows a similar result, on the full dataset the TFM performs well, and on the independent 2011 data set, it outperforms all other methods. Two criteria are used to evaluate the performance of the different methods. 1) The error percentage, which gives the percentage correctly predicted proteins, within a given error threshold, see eqn. (5.6). 2) The correlation between the predicted hydrophobic surface area and the hydrophobic surface area as calculated from the structure. The correlation is more sensitive to outliers when compared to the error threshold.

Table 5.1: Performance of TFM model compared to conventional, single amino acid predictors in the test set. The performance was measured in mean error (Å^2) and the Pearson correlation of the HSA determined with DSSP.

Tool	R	Mean error (Å^2)	R (2011)	Mean error (Å^2) (2011)
TFM	0.83	438.7	0.81	514.8
SANN	0.87	453.9	0.78	705
NetSurfP	0.82	558.5	0.75	668
SARPRED	0.77	599.5	0.76	822.2
SPINEX	0.78	707	0.76	754.6
Length derived	0.79	639.7	0.77	808.9

It should be stressed here that predicting the total hydrophobic surface area is a fundamentally different problem from predicting the amount of surface exposed area for individual amino acids. This may explain why a purpose built simple method can outperform a strategy that makes an explicit prediction for the accessible surface area for each individual residue.

TABLE 5.2: Performance of TFM model in the test set using only one or two features. The performance was measured in mean error (Å^2) and the Pearson correlation of the HSA determined with DSSP

Feature	Error (Å)	Correlation
Length	496	R=0.76
Hydrophobic	467.2	R=0.80
Hydrophilic	529.3	R=0.71
Length + Hydrophobic	448.7	R=0.83
Length + Hydrophilic	479.8	R=0.79
Hydrophobic+Hydrophilic	444.2	R=0.83

One of the benefits of the TFM method is that it is simple in terms of interpretation and implementation (see Figure 5.5S for the Python code implementation). Unlike the other methods, it is purely sequence based and does not rely on an evolutionary profile for its predictions. Furthermore, it requires little computational resources and is robust to overfitting since it has only three features (Figure 5.7S).

Interpretation of the three feature model

In order to understand why the three feature model works so well, we consider the three features used separately. We consider how well the method performs, when we include only one, two or all three features.

Table 5.2 shows the number of hydrophobic residues by itself is, unsurprisingly, the single strongest predictor for the total amount of hydrophobic surface area; adding the length improves the predictions further. We can understand this when we consider that proteins fold into a roughly globular structure, and are most stable when hydrophobic residues are buried inside the cores. Nevertheless, given the length of a protein chains, there is only a certain amount of residues that can be buried within the core, providing a physical restraint for the number of surface-exposed hydrophobic amino acids.

The predictions from the spherical model, which predicts the hydrophobic surface area from the length of the sequence alone are not accurate. This is not surprising, since the fraction of hydrophobic surface area, relative to the total surface area has a wide distribution in proteins, as shown in Figure 5.1S. Moreover, the ratio between hydrophobic and hydrophilic residues in the structure can explain part of the error (Figure 5.3S). Structures with a high ratio of hydrophobic over hydrophilic residues tend to have a higher HSA for a given sequence length. The TFM model is able to capture this information to make a more accurate prediction.

We suggest that the good performance of the full TFM method for hydrophobic surface prediction, can partially be explained by its ability to take

into take account the surplus of hydrophobic residues for a given length, i.e. hydrophobic residues that cannot be buried given the physically limiting surface over volume ratio.

Because the method relies on the assumption that proteins are approximately globular, it is possible that the accuracy of the predictions decreases for multiple chain proteins. However, we found that even for multiple chain proteins, the predictions are relatively accurate, see Figure 5.6S.

Improving predictions for per residue methods using TFM

Here we have shown that the features sequence length, the number of hydrophobic and hydrophilic amino acids can be used to predict the total hydrophobic surface area. Models for ASA prediction do not incorporate these global properties of proteins, and stay limited to local information using a windowed approach.

Here we explored if it is possible to improve such predictions, by taking into account the three global features as used in the TFM model. We trained an additional layer of the CUBIST algorithm, combining the ASA prediction from NetSurfP and features that were used in the TFM model. We find a small improvement in accuracy when we include these global features (Figure 5.4S). More specifically, the global features can indicate protein types that tend to under- or overpredict the total exposed (hydrophobic) surface area; this information can be used to correct per-residue predictions for the exposed surface area.

Conclusion

Above we have shown, show that a simple, rule-based method outperforms several machine learning based methods. While surprising, it has been found recently for other systems that simple methods can outperform more complicated models. For example, (Smialowski et al., 2007) show a machine-learning based method to predict the solubility of proteins expressed in *Escheria Coli*. Another example is given in (Vangone and Bonvin, 2015), where protein-protein interaction strength is predicted using a simple linear equation that only incorporates the non-interacting surface and the inter-residue contacts

Similarly, the TFM model outperforms adapted per residue methods in terms of predicting the hydrophobic surface area. The performance of the TFM model is not surprising because it was purposely build to predict the hydrophobic surface area. In contrast, the per-residue tools predict wether a residue is present on the surface. However, the performance increase when using the TFM model is significant and can potentially help in designing experimental protocols and interpreting results. Moreover, the performance of the per-residue tools can be increased by including global properties of the protein like the length of the sequence and number of hydrophobic and hydrophilic residues.

The model presented here shows that basic physical principles determine the total amount of hydrophobic surface. The performance of the simple model can be rationalized by the physical restraint on the volume of the core available for hydrophobic residues to be buried.

5.1 Supplementary Information - Predicting the hydrophobic surface area of native protein structures from sequence

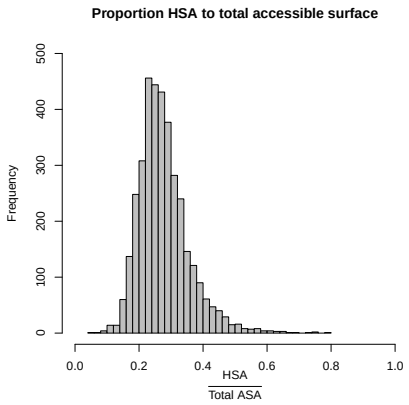

FIGURE 5.1S: Histogram of the total HSA over the total surface area. The average ratio is 0.28 with a 0.08 standard deviation.

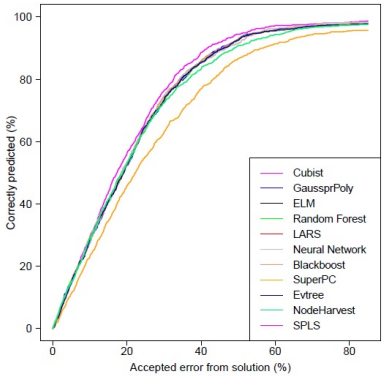

FIGURE 5.2S: The performance of several regression models, to include the three features length, number of hydrophobic and hydrophilic residues. The cubist algorithm slightly outperforms the other models considered.

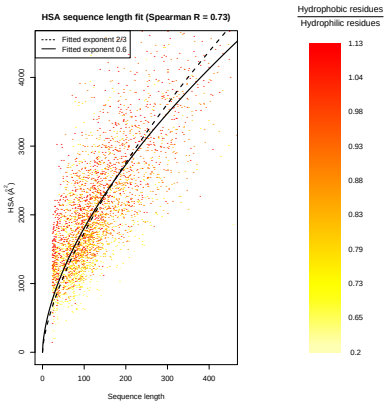

FIGURE 5.3S: Fitting the sequence length and the corresponding HSA. Perfect spheres with an equal ratio of exposed hydrophobic residues would follow the equation in eqn. (5.4). In this equation k_1 and k_2 were fitted at 116.5 and 0.60, respectively. Higher opacity of the data-points indicate a higher density at that position in the plot.

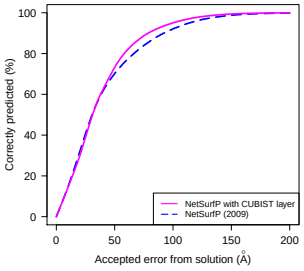

FIGURE 5.4S: Performance of NetSurfP with a extra layer of cubist with the features of TFM and the predicted ASA for individual amino acids. The correlation for individual amino acids improved from 0.33 to 0.39.

```
def cubistTFM(hydrophobic,hydrophilic,length):
    case = "C0"
    if hydrophobic <= 20:
        outcome = 592.2 + 79.1 * hydrophobic - (16.2 * hydrophilic)
        case = "C1"
    elif hydrophobic > 20 and  hydrophobic <= 99 and hydrophilic > 20:
        outcome = 1018.9 + 27.1 * hydrophobic - (7.6 * hydrophilic)
        case = "C2"
    elif hydrophobic > 20 and hydrophilic <= 20:
        outcome = 894.4 + 61.6 * hydrophobic - (0.3 * hydrophilic)
        case = "C3"
    elif hydrophobic > 99 and hydrophilic > 89:
        outcome = 934.9 + 27.2 * hydrophobic - (7.7 * hydrophilic)
        case = "C4"
    elif hydrophobic > 99 and hydrophilic <= 89:
        outcome = 1862.6 - 62.6 * length + 140 * hydrophobic
        case = "C5"
    return (outcome,case)
```

FIGURE 5.5S: Implementation of the TFM model in python.

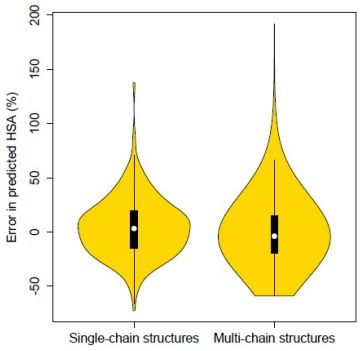

FIGURE 5.6S: Error in predicted HSA of proteins that consist of multiple chains or single chains.

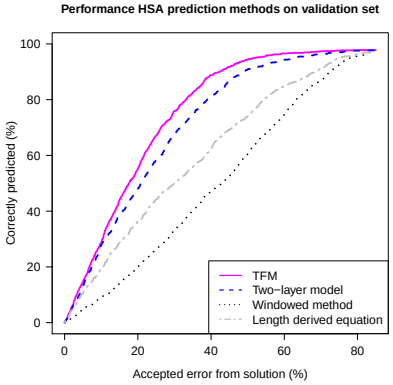

FIGURE 5.7S: Performance of the TFM model on the train, validation and test set.

TABLE 5.3: Calculated propensity values for an amino acid to be buried or exposed. The amino acid was considered exposed if more than 7 % of its total area was exposed.

Amino acid	Propensity buried
Ala	1.38
Cys	1.9
Asp	0.39
Glu	0.26
Phe	1.75
Gly	0.97
His	0.71
Ile	1.83
Lys	0.14
Leu	1.7
Met	1.48
Asn	0.52
Pro	0.66
Gln	0.4
Arg	0.31
Ser	0.8
Thr	0.86
Val	1.76
Trp	1.44
Tyr	1.16

Chapter 6

Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence

Based on the manuscript:

E van Dijk, H Mouhib, AK Buell, and S Abeln (2016a). "Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence". In: *In preparation*

Abstract

Proteins fold into a native shape to perform their function. Under some conditions, proteins can denature and aggregate together. The conditions under which proteins tend to aggregate are an active area of research. It is known to be influenced by PH, temperature, pressure and other factors. These aggregates can be either regular, structured beta-sheets, called amyloid fibres or disordered aggregates. Amyloid fibres are related to a number of degenerative diseases, amongst which are Alzheimer's disease and Parkinson's disease. Once fibres are formed, they are extremely stable and very resistant to denaturation by pressure or temperature. However, it has been reported that fibres can in rare cases denature at low temperature. We investigate the role of hydrophobicity in the stability of fibres. While it is known that almost all proteins can form amyloid fibres, hydrophobic sequences are more stable, suggesting a role of the hydrophobic effect. In this work we extend a published model with a consistent treatment of the temperature dependence of the hydrophobic effect. We show that this temperature dependence can reproduce cold denaturation of fibres, consistent with earlier experimental work.

Introduction

The folding, mis-folding, and aggregation of proteins are fundamental biological processes crucial to the functioning or disfunctioning of the cell; nevertheless, the physical models to describe these processes are still a topic of active research (Morriss-Andrews and Shea, 2015; Shakhnovich and Gutin, 1993b; Shakhnovich and Gutin, 1993a; van Dijk et al., 2016c; Shakhnovich, 1994; Sali, Shakhnovich, and Karplus, 1994; Coluzza, Muller, and Frenkel, 2003; Coluzza and Frenkel, 2004; Coluzza and Frenkel, 2007; Abeln and Frenkel, 2011). It is believed that the main driving force behind and protein folding processes in general is the hydrophobic effect (Baldwin, 2007). In general terms, the hydrophobic effect reflects the tendency of water to avoid immediate contact with non-polar particles such as hydrophobic amino acids or hydro-carbons. The effect arises mainly from the ability of water to form intermolecular hydrogen bond networks with itself. As a result, non-polar molecules tend to aggregate with each other and the hydrophobic effect provides the main driving force for the formation of biological self-assembled structures such as lipid membranes and proteins (Chandler, 2005; Dias et al., 2010). This is reproduced by several models for protein folding, where cold, heat, and pressure induced denaturation are studied including the effect of water as an explicit solvent (Sirovetz, Schafer, and Wolynes, 2015; Huang and Chandler, 2000; van Dijk et al., 2016c; Brotzakis et al., 2016). Much progress was also achieved in the study of protein denaturation and a recent study further confirms that alterations in the interactions between protein and water can be decisive in causing cold denaturation (Sirovetz, Schafer, and Wolynes, 2015). The hydrophobic effect is thus an important concept for the description and understanding of the underlying mechanisms of the fundamental biological process of protein folding.
In contrast to protein folding, there is currently no model that explicitly captures the hydrophobic effect, and especially its temperature dependence, in protein aggregation. Here, we use a model established for the study of aggregation (Abeln et al., 2014b; Ni et al., 2015) combined with a method to account for the temperature dependence of the hydrophobic effect used to study cold denaturation of protein folding (van Dijk et al., 2016c) to address this gap. We use this novel model to investigate the role of the temperature dependence of the hydrophobic effect in the thermodynamic stability of protein aggregates and explicitly investigate the elongation of an amyloid fibril.

The aggregation of peptides and proteins into regular amyloid fibrils is believed to be the underlying cause of many degenerative diseases. Even though many proteins have the potential to form amyloid fibrils under specific condition, diseases are often associated with specific proteins, e.g., fibrils of the Aβ peptide with Alzheimer's disease and α-synuclein fibrils with Parkinson's disease. Amyloid fibrils typically consist of beta-sheets and the aggregating regions are usually more hydrophobic than the surrounding regions. Under physiological conditions, amyloid fibrils are very stable (Baldwin et al., 2011), but a high barrier prevents aggregates from forming (Knowles et al., 2009; Buell

et al., 2010; Buell, 2011; Saric et al., 2016). At very high temperatures, amyloid fibrils can denature (Sasahara, Naiki, and Goto, 2005), and some fibrils can also cold denature at temperatures near the freezing point (Ikenoue et al., 2014).

Depending on the protein and the solution conditions there are multiple pathways from a monomeric protein to an amyloid fibril, consisting of different nucleation and growth processes. One of the challenges of the mechanistic studies of protein aggregation is to disentangle the overall aggregation process into the component elementary steps. This can be done either through global analysis of aggregation kinetics (Meisl et al., 2016) or else through measurements that are highly selective for a specific process. In the case of fibril elongation, this can be achieved in biosensing experiments, e.g. employing quartz crystal microbalances (QCM), whereby the growth of surface-bound seed fibrils upon incubation with monomeric protein is investigated (Knowles et al., 2009; Buell, Dobson, and Welland, 2012). Typically, proteins will form one or more oligomeric states before forming amyloid fibrils. However, this is not always the case. For example, light scattering experiments show that amyloid formation can proceed from the native state without an intermediate, oligomeric state or an unfolded state under certain conditions (Soldi et al., 2016). Additionally, some proteins form oligomers, but do not form amyloid fibrils (Gemma Soldi, Francesco Bemporad, and and Fabrizio Chiti*, 2008) and the amyloidogenic polypeptides α-synuclein and amyloid-beta can be redirected into oligomers by a small molecule, epigallocatechin gallate (EGCG) (Ehrnhoefer et al., 2008), demonstrating that amyloid fibrils are not under all conditions the thermodynamically most stable state. However, in the absence of any cofactors, amyloid fibrils are highly stable and the exact origin of this remarkable thermodynamic stability has not yet been elucidated.

In this study, we will focus on the thermodynamic signature associated with the elongation of a preformed amyloid fibril, without considering the kinetic or thermodynamic characteristics of the nucleation processes and the possible formation of oligomers as intermediate states before the formation of well-defined fibrils. The focus on the elongation step allows us to simplify the simulation setup, and to compare our simulations directly with experimental data available for this step.

Experimental thermodynamic studies show that amyloid fibrils are more temperature resistant than folded protein (most are stable up to at least 80 Centigrade), and that amyloid fibrils do not normally cold denature (e.g., (Sasahara, Naiki, and Goto, 2005; Morel, Varela, and Conejero-Lara, 2010; Kardos et al., 2004; Ikenoue et al., 2014)). The thermodynamic properties can be probed using Differential Scanning Calorimetry (DSC) or Isothermal Titration Calorimetry (ITC) (Jeppesen et al., 2010; Sasahara, Naiki, and Goto, 2005; Morel, Varela, and Conejero-Lara, 2010; Kardos et al., 2004). Under constant pressure, the heat (dQ) added or removed from the system is equal to its change in enthalpy (dH). Typically, ITC experiments show that aggregation is an exothermic process, and that the stability of aggregates decreases

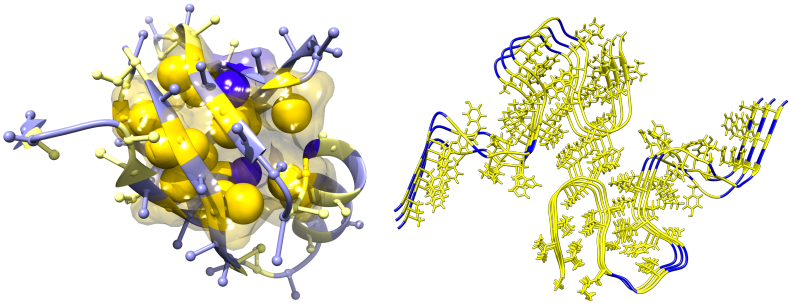

FIGURE 6.1: **Aggregated state and folded state**. Two alternative states for proteins, the functional native state of an iron transporter protein (PDB-ID:2K5I, left) and the disease-related amyloid sheet of the Aβ (1-42) peptide (PDB-ID: 2NAO) (Wälti et al., 2016) (right). Hydrophobic residues are coloured in yellow, while hydrophilic residues are coloured in blue. In a typical protein, a hydrophobic core can be observed, where the hydrophobic residues are shielded from the water by the hydrophilic residues. Aggregating proteins, and especially the core aggregating region of proteins tend to be more hydrophobic than normal proteins. The lack of protection from hydrophilic amino acids can cause a runaway aggregation of proteins (see left side).

with increasing temperature (Jeppesen et al., 2010; Sasahara, Naiki, and Goto, 2005; Morel, Varela, and Conejero-Lara, 2010; Kardos et al., 2004). However, a recent study showed that the aggregation of partial-length α-synuclein is an exception to the rule: it is an endothermic process, and its aggregates denature at low temperatures (Ikenoue et al., 2014; Kim et al., 2008).

In order to elucidate the molecular mechanisms behind the stability of the fibrils and protein aggregation, insight from computational models is required. Understanding the fundamental mechanisms behind aggregation potentially allows us to predict whether a sequence in certain conditions will aggregate. At this stage, coarse-grained simulations are the method of choice to study large systems with multiple protein chains. In addition to their low computational cost when compared to full-atomistic models, coarse grained model permit to isolate the minimal amount of necessary ingredients to understand a targeted molecular process. Here, we use a coarse-grained model to study the effect of temperature dependent hydrophobicity in protein aggregation.The present model is an extension of protein lattice models developed to study protein folding (Miyazawa and Jernigan, 1985b; Shakhnovich and Gutin, 1993b; Shakhnovich and Gutin, 1993a; van Dijk et al., 2016c; Shakhnovich, 1994; Sali, Shakhnovich, and Karplus, 1994; Coluzza, Muller, and Frenkel, 2003; Coluzza and Frenkel, 2004; Coluzza and Frenkel, 2007; Abeln and Frenkel, 2011). The

extended model includes an explicit representation and energy term for backbone hydrogen bonding (Abeln et al., 2014b) to allow for beta strand formation, that was used to study protein folding and aggregation (Ni et al., 2013b; Ni et al., 2015; Tran, Nguyen, and Derreumaux, 2016).

Recently, we have added a temperature dependence to the interactions of hydrophobic amino acids with water in a more general representation of the protein lattice model (Abeln and Frenkel, 2011), which allowed us to investigate the temperature dependence of the hydrophobic effect on the thermodynamic properties of protein folding (van Dijk et al., 2016c). In the model presented here, we incorporate both the hydrogen bonds and the temperature dependence of the hydrophobic effect in the lattice model as described in ref (Abeln et al., 2014b). Additionally, we introduce a new parameter, N_β, to model the entropic favourability of the formation of a beta sheet. The parameters in our model correspond to physical features of amyloid fibrils, i.e., the strength of the hydrogen bond interactions, the role of the hydrophobic effect in assembly, and the hydrophobicity of the sequence. This allows us to probe how these features determine the thermodynamic properties of aggregation.

Here we compare our model to several experimental observations on the thermodynamics characteristics of amyloid fibrils: 1) Under normal conditions (i.e., pH 5-10, intermediate temperatures, no denaturant) the fibrillar state is thermodynamically stable; moreover, amyloid formation is typically an exothermic process, except in some rare cases (Ikenoue et al., 2014). 2) At very high temperatures, most amyloid fibrils denature and the fibrils become soluble. 3) For aggregates that are marginally stable under normal conditions (see above), fibrils can become soluble at low temperatures, which leads to cold denaturation of the fibrils through dissociation into unstructured protein monomers(Ikenoue et al., 2014). Most of these observations can indeed be reproduced with our model, for ranges of model parameters that are physically plausible. Our model also predicts that the slope of ΔH for amyloid formation with respect to temperature is directly related to the hydrophobic effect.

Methods

Experimentally, it has been shown that aggregation is often strongly accelerated in the presence of preformed fibrillar aggregates, or seeds. In this study, we focus on the thermodynamic properties of elongation, or the addition of a single protein molecule to the end of a 'seed' fibril. We study this process using a cubic lattice model, where each amino acid occupies a single grid site. In this model, inspired by the model introduced by Shakhnovic (Shakhnovich and Gutin, 1993b; Shakhnovich and Gutin, 1993a; Shakhnovich, 1994; Sali, Shakhnovich, and Karplus, 1994; Coluzza, Muller, and Frenkel, 2003; Coluzza and Frenkel, 2004; Coluzza and Frenkel, 2007), each amino acid interacts with the amino acids directly adjacent to it. If no amino acid is present, the grid site is assumed to be occupied by the solvent (Abeln and Frenkel, 2011).The side chain is modelled by giving each amino acid a direction, and allowing hydrogen

bonds to be formed only when the side chains point towards the same direction, allowing a reasonable steric approximation of beta strands (Abeln et al., 2014b).

A full mathematical description of the model is given by the following equation:

$$\mathcal{H} = E_{\text{hb}} + E_{\text{steric}} + E_{\text{solvent}} + E_{\text{state}} + E_{a_i,a_j} + \Phi(T) \tag{6.1}$$

The term E_{a_i,a_j} represents the classical pairwise amino acid interactions, that are used in most coarse-grained simulations. The term E_{hb} represents the interactions originating from hydrogen bonds between side chains of two amino acids. Note that, since our simulations have an implicit solvent, ϵ_{hb} indicates the difference in energy between a hydrogen bond formed in bulk water and between amino acids. E_{solvent} is the interaction of an amino acid with the solvent. E_{state} represents the energy gained from β-sheet formation. The Hamiltonian, $\mathcal{H}$ is given in reduced units (k_BT units). The first four terms are kept identical to the methods described by Abeln et al. (Abeln et al., 2014b). Below we give a short summary of the model, starting with the addition to this model describing the hydrophobic effect.

Hydrophobic effect

The term $\Phi(T)$, which accounts for the temperature dependence of the hydrophobic effect was recently introduced by us in (van Dijk et al., 2016c). This term allows us to account for the temperature- and length scale dependence of the hydrophobic effect. Since there are hydrogen bonds present in our model, this term is no longer required. Instead, we have added the requirement that an amino acid which does not make a hydrogen bond with another amino acid to have an interaction with water. Thus, the mathematical form of $\Phi(T)$ is:

$$\Phi(T) = \Sigma_i F_{\text{hydr}} \tag{6.2}$$

where F_{hydr} is defined as:

$$F_{\text{hydr}} = \alpha(T - T_0)^2 \epsilon_{a_i,w} \tag{6.3}$$

Where α and T_0 control the strength of the hydrophobic effect, and $\epsilon_{a_i,w}$ indicates the interaction that amino acid i has with the solvent. The correction is only applied to hydrophobic amino acids. A hydrophobic amino acid is defined as an amino acid where $\epsilon_{a_i,w} > 0$. In our potential, the amino acids that fulfill this criterium are $a_i \in \{C, F, L, W, V, I, M, T, A\}$. In our experiments, we investigated the sensitivity of the model to the hydrophobic effect by investigating the cases $\alpha = 0$, $\alpha = 20$, $\alpha = 40$ and $\alpha = 60$.

Hydrogen bonds

The term describing hydrogen bonds can be written as $E_{\text{hb}} = \sum \epsilon_{hb} H_{i,j} \cdot C_{i,j}$, where ϵ_{hb} represents the potential energy per hydrogen bond. $H_{i,j} = 1$ indicates

that a hydrogen bond is present, and $H_{i,j} = 0$ indicates that no hydrogen bond is present. In our model, the water interactions are implicit, so ϵ_{hb} indicates the difference between a hydrogen bond of an amino acid with the solvent and a hydrogen bond in bulk water. Hydrogen bonds between amino acids and the solvent are typically stronger than the hydrogen bonds in a bulk solvent. We investigate the cases $\epsilon_{\mathrm{hb}} \in \{0.25, 0.5, 0.75, 1.1, 5\}$

Entropically favorable β-sheets

We also investigated the effect of an entropical bonus for β-sheets. The high stability of β-sheets at elevated temperatures suggests that β-sheets have a higher entropy than other secondary structure elements. This could be due to a large degree of flexibility of the side chains in a β-sheet. We investigated this hypothesis by introducing a number of degenerate states, N_β if the state of an amino acid is a β-sheet. These degenerate states are identical to the normal β-sheet state, but allow us to investigate the effect of an entropic 'bonus' an amino acid receives for being in a β-sheet. Unless otherwise stated, $N_\beta = 1$, and we investigated the case for $N_\beta = 5$ (Figure 6.2S)

Model setup

In our simulations, a seed is represented by a preformed fibril consisting of 8 peptides. This fibril is 'frozen' during our simulations, which means that the amino acids in this seed are not allowed to make any moves, aside from moves that leave the overall structure intact: moves that translate the entire seed fibril, and moves that rotate the entire seed fibril. However, all interactions of the fibril with the environment are still present. Two additional monomeric protein molecules are present in the simulation box, which are allowed to make regular moves, and can attach and detach from the seed during the simulation. This setup allows us to investigate the addition of one layer (consisting of two molecules) to the preformed seed fibril.

Evaluation

We used the umbrella sampling method to sample the conformational space. Umbrella sampling biases the simulation across an order parameter, after which the free energy landscape can be obtained with the Weighted Histogram Analysis Method (WHAM) (Grossfield, 2003). As order parameter, we used the number of external contacts, C_{ext}. We use a quadratic biasing potential to define E_{umbr}

$$E_{\mathrm{umbr}} = \mathcal{H} + k(C_{\mathrm{ext}} - C_0)^2 \tag{6.4}$$

Where k is the spring constant, $\mathcal{H}$ the Hamiltonian defined in eqn. (6.1), and $C_{\mathrm{ext},0}$ the value towards which the simulation is biased. In our simulations, $k = 2$ and $C_{\mathrm{ext},0} \in \{70, 75, 80, 85, 90, 95\}$

Definition of aggregated state

An aggregate is defined to be in the fibrillar state when at least 88 out of the maximum 91 external contacts are present.

$$\text{Aggregate} = \begin{cases} 1 \text{ if } C_{\text{ext}} >= 88 \\ 0 \text{ if } C_{\text{ext}} < 88 \end{cases} \tag{6.5}$$

We define $< P_{\text{Aggregate}} >$, as shown in Figure 6.3 as the ensemble average of the "Aggregate" observable.

Sequences

For the peptides we used the same sequences as in ref. (Abeln et al., 2014b): TFTFTFTF. To investigate the effect of a lower stability, we replaced the Phenylalanine residues with Leucine residues.

Results and discussion

Structural features of the cold denatured state

First, we investigated the addition of the hydrophobic effect and its effect on the stability of the fibril. We used the same fibrils as used in refs. (Abeln et al., 2014b; Ni et al., 2013b). In this previous work, the temperature dependence of the hydrophobic effect was not taken into account. We used the parameter α to set the strength of the hydrophobic effect, as in Ref. (van Dijk et al., 2016c). For $\alpha = 40$, we observe both cold and heat denaturation of the fibril as shown in Figure 6.2.

Experimentally, it appears that the stability of a fibril at physiological temperatures is strongly correlated to the stability at low temperatures. In the model described here, we can adjust the stability of the fibril by changing the strength of the hydrogen bonds. In the case of folding, the cold denatured state is thermodynamically distinct from the heat denatured state. It has also been observed both experimentally and in simulations that the cold unfolded state of a protein adopts a residual, non-native structure (Vajpai et al., 2013; van Dijk et al., 2016c). In our simulations, we find that the cold denatured ensemble displays less hydrophobic interactions (See Figure 6.2 and 6.3).

Experimental work suggests that cold denaturation of amyloid fibrils may occur in an analogous fashion. For some aggregates, oligomeric states can be found at low temperatures, as opposed to completely soluble, monomeric protein molecules that are found at very high temperatures (Ikenoue et al., 2014). This is in agreement with our simulations, where the heat denatured state consists of completely desolved fibrils, while the cold denatured state exhibits residual contacts, albeit less ordered (Figure 6.2.

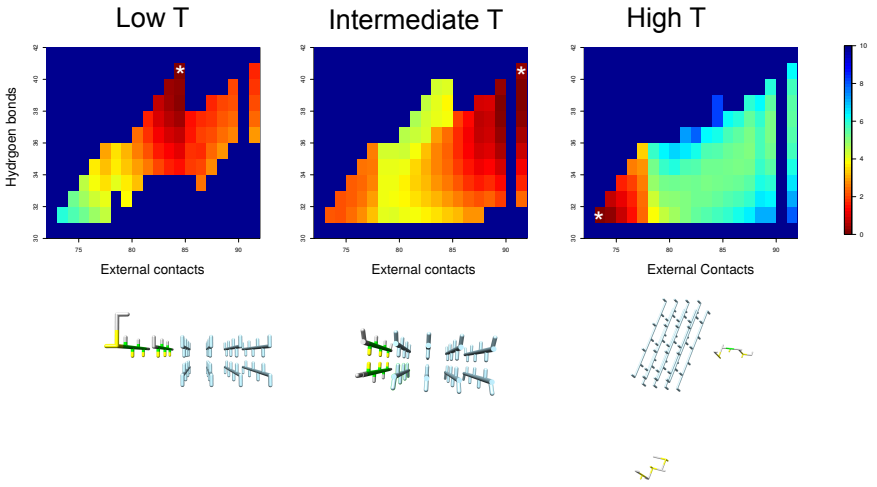

FIGURE 6.2: **Cold denaturation heatmap** Free energy landscapes and representative snapshots for each temperature, using $\alpha = 40$. The conformation with the lowest free energy is the most represented conformation in our simulation, as indicated by stars in the energy landscape. Note that the simulations were run with a preformed, frozen seed (indicated in light blue in the snap shots). At low temperatures, the heatmap shows that the simulated monomers adopt a conformation that has many 'native' contacts and H-bonds, but does not allow for unlimited, regular growth. Experimental results also show that cold denaturation can lead to the formation of oligomers (Kim et al., 2008). At intermediate temperatures, the monomers adopt a regular, fibrillar structure at the end of the seed fibrils. At high temperatures, only transient contacts are formed and the protein molecules that are not part of the seed remain fully solvated.

For real proteins, the strength of hydrogen bonds varies depending on the pH of the solution, the type of contributing amino acids and the global constraints imposed by the configuration of the fibril. Moreover, other factors that stabilise proteins through enthalpic interactions, salt bridges or van der Waals interactions should follow the same qualitative pattern. In our simulations, the stability of aggregates at low temperatures depends on the stability at physiological temperatures (see Figure 6.3). It must be noted that, for most types of fibrils, cold denaturation does not occur above freezing temperatures (Ikenoue et al., 2014). In our model, for realistic settings of α (eg. $\alpha \leq 20$, see Ref. (van Dijk et al., 2016c)) and the strength of a hydrogen bond, ϵ_{hb} (between 0.5 and 1 in kT units) no cold denaturation occurs at temperatures around the freezing point (corresponding to approximately T = 0.18 in reduced units).

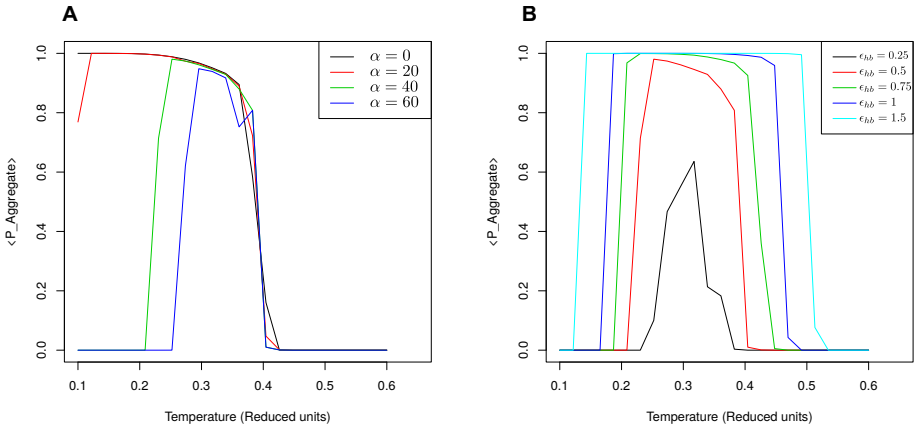

FIGURE 6.3: **Stability as explored for different parameter settings** Here, we explored the effect of different values of α on the stability of the aggregates (a) and the stability for different strengths of the H-bonds, controlled by the parameter ϵ_{hb} (b)

For sequences that are marginally stable (sequences with a weaker pairwise amino acid interaction, see Figure 6.1S, or a weaker H-bond, see Figure 6.3 (b)) the aggregates do denature at low temperatures. An example is given in Figure 6.1S, where the Phenylalanine amino acids have been replaced by Leucine amino acids, which decreases the stability. <P_Aggregate> peaks at 98% for this sequence.

Thermodynamics of aggregation

The thermodynamics of fibril formation is difficult to study experimentally, since aggregation is typically an irreversible process, rendering the application of some equilibrium thermodynamic methods, such as DSC, problematic. However, if the enthalpy difference between two states is temperature independent, a system will adopt an enthalpically favourable state at low temperatures, and an entropically favorable state at high temperatures. Combining this with the fact that most amyloid fibrils become soluble only at high temperatures suggests that amyloid fibrils are enthalpically stable. This is supported by most calorimetry (DSC and ITC) experiments (Kardos et al., 2004; Morel, Varela, and Conejero-Lara, 2010; Ikenoue et al., 2014).

We investigated if these results hold in our model, for different parameters. For the temperature-independent potential ($\alpha = 0$), the fibrils are the enthalpically stable state (corresponding to Δ H < 0 in Figure 6.4. For the temperature dependent potential, this depends on the parameters in the model. In Figure 6.4, the enthalpy of aggregation is shown for various parameter settings. In the model described here, we can adjust the stability of the fibril by changing

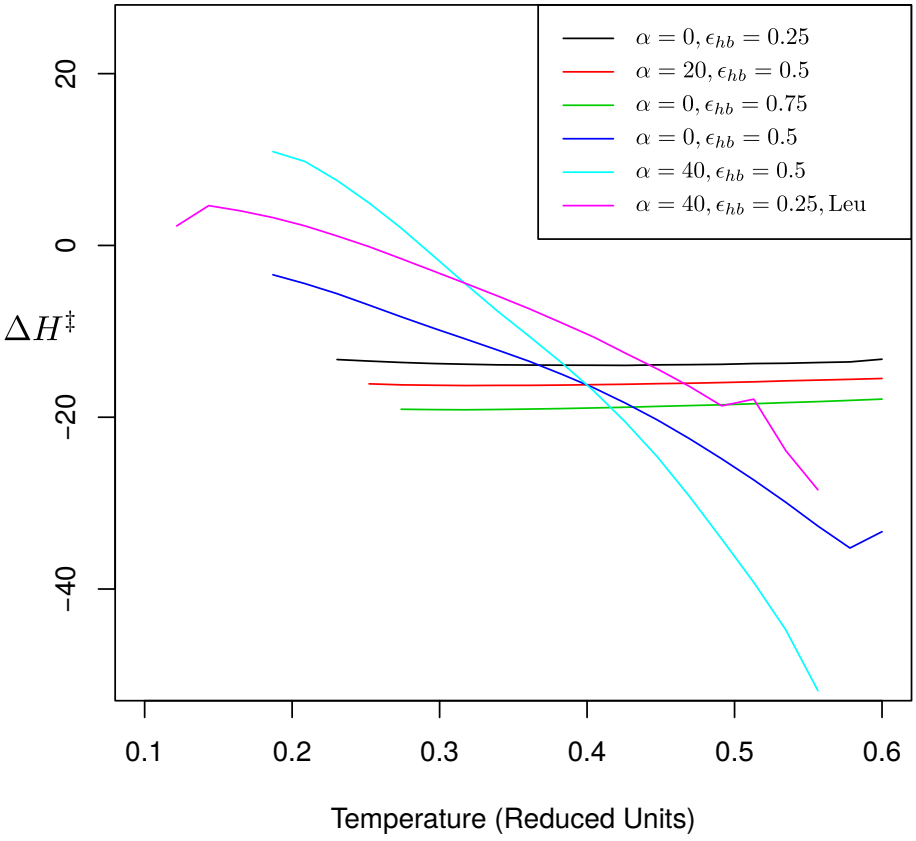

FIGURE 6.4: **Enthalpy of aggregation** In our model, the slope of the enthalpy of fibril elongation as a function of temperature, the heat capacity of the elongation reaction, is determined by the temperature dependence of the hydrophobic effect (α). Increased strength of the H-bonds simply increase the difference in enthalpy between the fibril and the desolved state. This effect directly contributes to the stability of the fibril.

the strength of the hydrogen bonds. The stability obtained through H-bonds is purely enthalpic, because changing the strength of H-bonds simply modifies the enthalpy of aggregation. The stability obtained through these H-bonds is additive in our model.

The temperature independent potential leads to a constant enthalpy of aggregation, independent of temperature. Adjusting the strength of the hydrophobic effect leads to a slope in the enthalpy as a function of temperature. A negative slope can be found for most amyloid fibrils in experimental work as well (Kardos et al., 2004; Jeppesen et al., 2010; Ikenoue et al., 2014). However,

in (Ikenoue et al., 2014) it is reported that α-synuclein has a small positive heat capacity of fibril elongation, as well as being entropically favorable at all temperatures. The fibrils that denature at low temperatures in our model are enthalpically favorable at high temperatures, but entropically favourable at low temperatures. See the light blue and magenta lines in Figure 6.3 (b).

Conclusion

In this work, we investigate how physical parameters affect the behavior of aggregating proteins after a nucleus has been formed. We show that realistic settings of these physical parameters, comparable to settings that are required to obtain self-replication in a simpler model (Saric et al., 2016), allow the model to reproduce the behaviours that fibrils show under changing conditions, specifically: 1) The model prediction that under normal conditions the fibrillar state is thermodynamically stable, as experimentally observed for α-synuclein, β2 microglobin, insulin fibrils, the K3-fibrils and the three types of fibrils of amyloid β(Aβ)1-42 and Aβ1-40 peptides (Normal conditions are defined here as: pH 5-10, intermediate temperatures and no denaturant). Moreover, the model shows that fibril formation is typically an exothermic process. Experimentally the same observation has been made, except under some very specific conditions for α-synuclein (Ikenoue et al., 2014; Kim et al., 2008) 2) The model suggest that at very high temperatures, amyloid fibrils denature and the fibrils become soluble; this can also be observed experimentally. 3) The model shows that for aggregates that are marginally stable under normal conditions fibrils can become soluble at low temperatures, i.e. denaturation of the fibrillar state; as has been observed for α-synuclein (Ikenoue et al., 2014)) The model also suggest that slope of ΔH with respect to the temperature for the process of amyloid formation is directly related to the temperature dependence of the hydrophobic effect.

Acknowledgments

SA has been supported by a Veni grant on the project 'Understanding toxic protein oligomers through ensemble characteristics' from Netherlands Organisation for Scientic Research (NWO).

6.1 Supplemental information - Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence

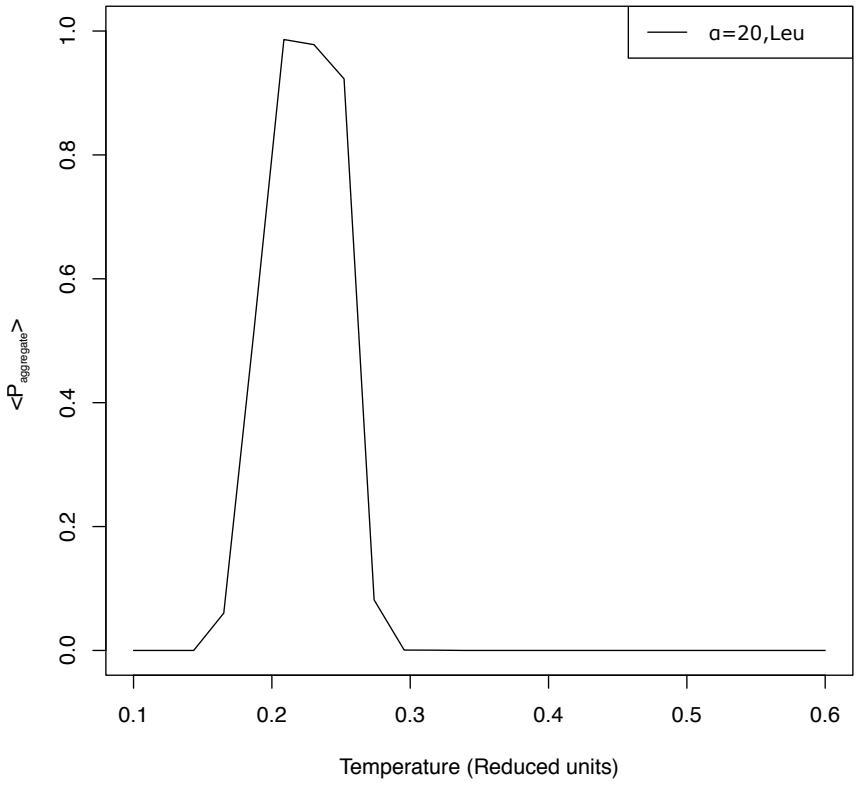

FIGURE 6.1S: **Stability Leu** The stability of a peptide with sequence TLTLTLTL. Leucine has weaker pairwise interactions with itself, decreasing the stability of the fibril. This causes the fibril to denature at low temperatures.

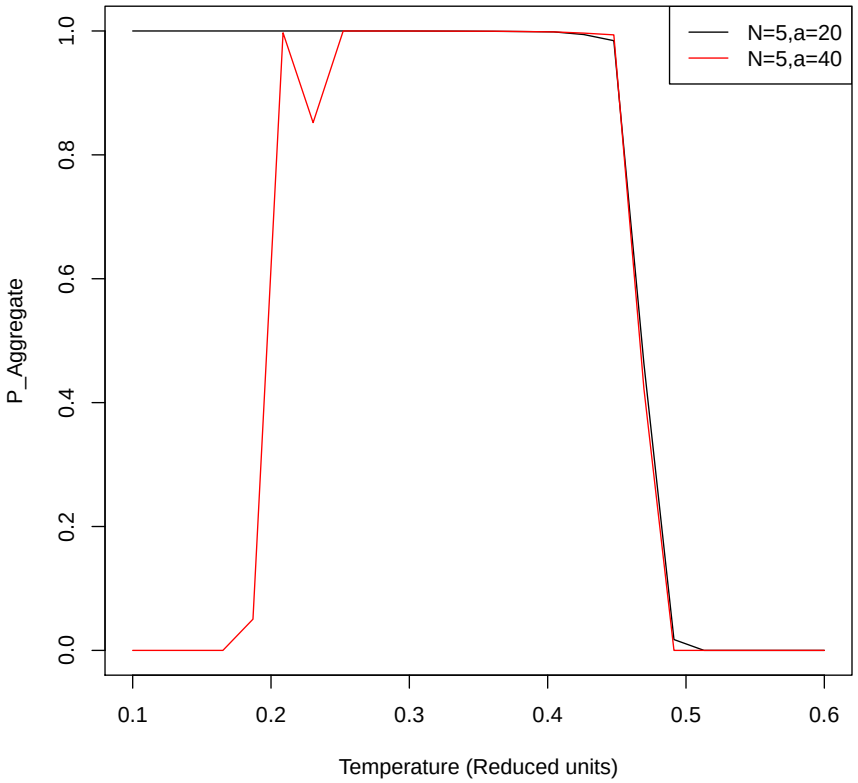

FIGURE 6.2S: **Stability Num states** The stability of a peptide with sequence TFTFTFTF, where the amino acids in a strand have more degenerate states. Incorporating this effect causes the fibrils to have a higher stability over all temperatures.

Chapter 7

Discussion

The importance of the hydrophobic effect in protein folding and aggregation has been textbook knowledge for decades. For example, the undergraduate textbook "Introduction to protein structure" (Branden and Tooze, 1998), starts with the observation that the interior of proteins is hydrophobic. This observation is than followed by the statement: "The main driving force for folding water-soluble globular protein molecules is to pack hydrophobic side chains into the interior of the molecule, thus creating a **hydrophobic core** and a hydrophillic surface".

Moreover, Spolar et al. showed that important thermodynamic variables, the entropy and enthalpy of folding, and the related experimental observable, the heat capacity of the system, were determined by the interactions of the hydrophobic amino acids with the water (Spolar, Ha, and Record, 1989). Nevertheless, quantifying the role of the *temperature dependence* of the hydrophobic effect in protein folding, and in particular its contribution to cold denaturation and to the temperature dependence of the heat capacity, were studied in less detail.

In contrast to protein folding, little was known on the role played by hydrophobic contributions to aggregating proteins at the start of this project. Experimental results and theory showed that transforming a protein to a fibril occurred in three stages: oligomerization, nucleation, and elongation (Knowles et al., 2009). Knowles et al. also showed that different sequences behaved differently, and that the mechanism of aggregation could be determined by fitting a kinetic model to the experimentally observed rate of aggregation. This approach has gained wider adoption and has now been published as a method (Meisl et al., 2016). This approach does not provide insight on the molecular pathway from a folded protein to an aggregated fibril, nor on the final structure of the amyloid fibril. The leading hypothesis during this period was was that the structures of the final, fibrillar state consisted of twisted β-sheets. This hypothesis was strengthened by the fact that coarse grained computational models could reproduce the aggregated state, and the elongation process that allowed aggregates to grow. However, no models capture the trade-off between folding, oligomer formation, and elongation that exists in real proteins.

In this thesis, we have shown that cold denaturation, the heat capacity and the structure of the cold denatured state can be explained by incorporating the

temperature dependence of the hydrophobic effect. The coarse grained nature of our model allows us to find general principles and statistical correlations that can be applied to real proteins. We made a result from our model, prediction of the heat capacity of a given folded protein from the hydrophobic surface area, publicly accessible through a web server. Ultimately one would like to use simulations to make accurate, physics-based predictions. Two publications published concurrently with our work show that the general principles we found are also applicable to make quantitative predictions for the stability of real proteins (Sirovetz, Schafer, and Wolynes, 2015; Vajpai et al., 2013). I expect that in the near future these methods will be applied to more proteins to validate the results further. Additionally, these methods can likely be applied to predict protein stability as a function of temperature.

While much progress has been made during this project on the understanding of protein aggregation, there are still a lot of unknowns. In our computational model we used the commonly accepted hypothesis that amyloid fibrils consist of twisted β-sheets. This year a structure was found for α-synuclein that was consistent with this hypothesis. During my PhD, a paper describing a model that could model the trade off between folding, oligomer formation and aggregation was published by (Ni et al., 2015). We have used their model in combination with the temperature dependent hydrophobic effect described in chapter 3 to investigate if aggregation was enthalpically or entropically driven. We also investigated the role of the temperature dependent hydrophobic effect in the temperature dependence of the enthalpy and the entropy. Unlike in protein folding, our results agreed only partially with published experimental work. While our model did reproduce cold denaturation of aggregates, a phenomenon that was first observed during my PhD, we found that the slope of the enthalpy of aggregation with regard to temperature was always smaller than zero, irregardless of the parameters we choose. Unpublished data from an experimental collaborator showed that the experimental result we were comparing to could not be reproduced in his lab. This is an interesting result that will be investigated further in a follow-up project.

While quantitative predictions for the stability of single protein structures are constantly being evaluated and improved, for example through the CASP project, the predictions made so far for aggregating proteins are mostly qualitative. An important reason for this contrast is that the computational resources required to simulate such a big system in atomic detail are enormous. This problem previously existed for protein folding and was solved by the increase in computational power due to Moore's law, which states: the number of transistors in a dense integrated circuit doubles approximately every two years. While Moore's law still holds approximately, it seems to no longer translate directly to increased processing power. A breakthrough in algorithms, model abstraction, or in computer hardware (and possibly two or more of the above) is required to allow us to accurately predict the conditions under which a sequence will aggregate.

I hope that this thesis provides a starting point towards the goals outlined

above. To allow our work to be continued, we have attempted to make our research reproducible: We made the programs we developed open source and included the data we used to generate our results. Moreover, we also made the work that is directly applicable to experiments accessible through a web server.

Table 7.1: Table of scientific software and links

Name	Ref.	Code
Hydrophold	(van Dijk et al., 2016c)	`https://bitbucket.org/edk360/hydrophold`
StrandcoilHydrophobe	(van Dijk et al., 2016a)	`https://bitbucket.org/edk360/strandcoil_hydrophobe`
HSAPred	(Bouwmeester et al., 2016)	`https://bitbucket.org/Robbinbwm/hsa`
BICEP	(van Dijk et al., 2016b)	`https://bitbucket.org/Robbinbwm/bicep`
Data hydrophobicity	(van Dijk, Hoogeveen, and Abeln, 2015)	`goo.gl/uRsgXN`

List of Publications

[1] R Bouwmeester, E van Dijk, J Heringa, and S Abeln. Predicting the hydrophobic surface area of native protein structures from sequence. *In preparation*, 2016.

[2] Ali May, René Pool, Erik van Dijk, Jochem Bijlard, Sanne Abeln, Jaap Heringa, and K Anton Feenstra. Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics*, 30(3):326–334, 2014.

[3] Margriet M Palm, Marchien G Dallinga, Erik van Dijk, Ingeborg Klaassen, Reinier O Schlingemann, and Roeland M H Merks. Computational Screening of Tip and Stalk Cell Behavior Proposes a Role for Apelin Signaling in Sprout Progression. *PLOS ONE*, 11(11):1–31, 2016.

[4] E van Dijk, R Bouwmeester, BJ Brandt, J Heringa, and S Abeln. Heat Capacity Baseline Prediction Using BICEP. *In preparation*, 2016.

[5] E van Dijk, H Mouhib, AK Buell, and S Abeln. Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence. *In preparation*, 2016.

[6] Erik van Dijk, Arlo Hoogeveen, and Sanne Abeln. The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLoS Comput Biol*, 11(5):e1004277, 2015.

[7] Erik van Dijk, Patrick Varilly, Tuomas P J Knowles, Daan Frenkel, and Sanne Abeln. Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Phys. Rev. Lett.*, 116(7):78101, feb 2016.

Bibliography

Abagyan, R A and S Batalov (1997). "Do aligned sequences share the same fold?" In: *Journal of molecular biology* 273.1, pp. 355–68. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1287. URL: http://www.sciencedirect.com/science/article/pii/S0022283697912870.

Abeln, S (2007). "Protein fold evolution on completed genomes : distinguishing between young and old folds". PhD thesis. URL: http://ora.ouls.ox.ac.uk/objects/uuid:b520fd65-e829-4ae0-bed6-47d642909889.

Abeln, Sanne and Daan Frenkel (2008). "Disordered Flanks Prevent Peptide Aggregation". In: *PLoS Comput Biol* 4.12, e1000241. DOI: 10.1371/journal.pcbi.1000241. URL: http://dx.doi.org/10.1371/journal.pcbi.1000241.

— (2011). "Accounting for Protein-Solvent Contacts Facilitates Design of Nonaggregating Lattice Proteins". In: *Biophysical Journal* 100.3, pp. 693–700. ISSN: 0006-3495. DOI: DOI:10.1016/j.bpj.2010.11.088. URL: http://www.sciencedirect.com/science/article/B94RW-52301WF-13/2/cc7e3cdcc25f0fb2533fda30a665dd18.

Abeln, Sanne et al. (2014a). "A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation". In: *PLoS One* 9.1. Ed. by Ilia V. Baskakov, e85185. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0085185. URL: http://dx.plos.org/10.1371/journal.pone.0085185.

Abeln, Sanne et al. (2014b). "A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation". In: *PLoS ONE* 9.1, e85185. DOI: 10.1371/journal.pone.0085185. URL: http://dx.doi.org/10.1371\%2Fjournal.pone.0085185.

Anslyn, Eric V. and Dennis A. Dougherty (2006). "Solutions and Non-Covalent binding forces". In: *Modern Physical Organic Chemistry*. University Science. Chap. 3, pp. 145–206. ISBN: 978-1-891389-31-3.

Auer, Stefan et al. (2008). "A Generic Mechanism of Emergence of Amyloid Protofilaments from Disordered Oligomeric Aggregates". In: *PLoS Comput Biol* 4.11, e1000222. DOI: 10.1371/journal.pcbi.1000222. URL: http://dx.doi.org/10.1371/journal.pcbi.1000222.

Baldwin, Andrew J et al. (2011). "Metastability of Native Proteins and the Phenomenon of Amyloid Formation". In: *Journal of the American Chemical Society* 133.36, pp. 14160–14163. ISSN: 0002-7863. DOI: 10.1021/ja2017703. URL: http://dx.doi.org/10.1021/ja2017703.

Baldwin, Robert L (2007). "Energetics of Protein Folding". In: *J. Mol. Biol.* 371.2, pp. 283–301. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2007.05.078. URL: http://www.sciencedirect.com/science/article/pii/S0022283607007371.

Berman, H M et al. (2000a). "The Protein Data Bank". In: *Nucleic Acids Res* 28.1, pp. 235–242.

Berman, Helen M et al. (2000b). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242. DOI: 10.1093/nar/28.1.235. URL: http://nar.oxfordjournals.org/content/28/1/235.abstract.

Betancourt, M R and D Thirumalai (1999). "Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes". In: *Protein Sci* 8.2, pp. 361–369. ISSN: 0961-8368. DOI: 10.1110/ps.8.2.361.

Beyreuther, K et al. (1991). "Mechanisms of amyloid deposition in Alzheimer's disease." en. In: *Annals of the New York Academy of Sciences* 640, pp. 129–39. ISSN: 0077-8923. URL: http://europepmc.org/abstract/med/1776729.

Bianco, Valentino and Giancarlo Franzese (2015). "Contribution of Water to Pressure and Cold Denaturation of Proteins". In: *Phys. Rev. Lett.* 115.10, p. 108101. DOI: 10.1103/

PhysRevLett.115.108101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.115.108101.

Blundell, T L et al. (1971). "Atomic Positions in Rhombohedral 2-Zinc Insulin Crystals". In: *Nature* 231.5304, pp. 506–511. URL: http://dx.doi.org/10.1038/231506a0.

Boswell, Paul G et al. (2011). "Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles". In: *Journal of Chromatography A* 1218.38, pp. 6742–6749. ISSN: 0021-9673. DOI: http://dx.doi.org/10.1016/j.chroma.2011.07.070. URL: http://www.sciencedirect.com/science/article/pii/S0021967311011095.

Bouwmeester, R et al. (2016). "Predicting the hydrophobic surface area of native protein structures from sequence". In: *In preparation*.

Branden, Carl and John Tooze (1998). *No Title*. 2nd ed. New York: Garland Publishing, p. 14.

Branden, Carl Ivar and JohnTooze (1999). *Introduction to Protein Structure*. ISBN: 1136969896. URL: https://books.google.com/books?hl=nl{\&}lr={\&}id=eUYWBAAAQBAJ{\&}pgis=1.

Brotzakis, Z F et al. (2016). "Dynamics of Hydration Water around Native and Misfolded α-Lactalbumin". In: *The Journal of Physical Chemistry B* 120.21, pp. 4756–4766. DOI: 10.1021/acs.jpcb.6b02592. URL: http://dx.doi.org/10.1021/acs.jpcb.6b02592.

Bruscolini, Pierpaolo and Athi N Naganathan (2011). "Quantitative Prediction of Protein Folding Behaviors from a Simple Statistical Model". In: *Journal of the American Chemical Society* 133.14, pp. 5372–5379. DOI: 10.1021/ja110884m. URL: http://pubs.acs.org/doi/abs/10.1021/ja110884m.

Buchete, N-V, J E Straub, and D Thirumalai (2004). "Development of novel statistical potentials for protein fold recognition". In: *Curr. Opin. Struct. Biol.* 14.2, pp. 225–232. ISSN: 0959-440X. DOI: http://dx.doi.org/10.1016/j.sbi.2004.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0959440X04000351.

Buell, Alexander K (2011). "On the Kinetics of Protein Misfolding and Aggregation". PhD thesis. University of Cambridge.

Buell, Alexander K, Christopher M Dobson, and Mark E Welland (2012). "Measuring the Kinetics of Amyloid Fibril Elongation Using Quartz Crystal Microbalances". In: *Amyloid Proteins: Methods and Protocols*. Ed. by M Einar Sigurdsson, Miguel Calero, and María Gasset. Totowa, NJ: Humana Press, pp. 101–119. ISBN: 978-1-61779-551-0. DOI: 10.1007/978-1-61779-551-0_8. URL: http://dx.doi.org/10.1007/978-1-61779-551-0{_}8.

Buell, Alexander K et al. (2010). "Frequency Factors in a Landscape Model of Filamentous Protein Aggregation". In: *Phys. Rev. Lett.* 104.22, p. 228101. DOI: 10.1103/PhysRevLett.104.228101.

Chamberlin, Adam C, Christopher J Cramer, and Donald G Truhlar (2006). "Predicting Aqueous Free Energies of Solvation as Functions of Temperature". In: *The Journal of Physical Chemistry B* 110.11, pp. 5665–5675. DOI: 10.1021/jp057264y. URL: http://dx.doi.org/10.1021/jp057264y.

Chandler, David (2005). "Interfaces and the driving force of hydrophobic assembly". In: *Nature* 437.7059, pp. 640–647. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature04162.

Chothia, C (1976a). "The nature of the accessible and buried surfaces in proteins." In: *J. Mol. Biol.* 105.1, pp. 1–12. ISSN: 0022-2836. URL: http://www.ncbi.nlm.nih.gov/pubmed/994183.

Chothia, C and J Janin (1975). "Principles of protein-protein recognition." In: *Nature* 256.5520, pp. 705–8. ISSN: 0028-0836. URL: http://www.ncbi.nlm.nih.gov/pubmed/1153006.

Chothia, Cyrus (1976b). "The nature of the accessible and buried surfaces in proteins". In: *Journal of Molecular Biology* 105.1, pp. 1–12. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/0022-2836(76)90191-1. URL: http://www.sciencedirect.com/science/article/pii/0022283676901911.

Cilia, Elisa et al. (2013). "From protein sequence to dynamics and disorder with DynaMine". In: *Nat Commun* 4. URL: http://dx.doi.org/10.1038/ncomms374110.1038/ncomms3741.

Coluzza, I, H G Muller, and D Frenkel (2003). "Designing refoldable model molecules". In: *Phys. Rev. E* 68.4, p. 46703. DOI: 10.1103/PhysRevE.68.046703.

Coluzza, Ivan (2011). "A Coarse-Grained Approach to Protein Design: Learning from Design to Understand Folding". In: *PLoS ONE* 6.7, e20853. DOI: 10.1371/journal.pone.0020853. URL: http://dx.doi.org/10.1371/journal.pone.0020853.

Coluzza, Ivan and Daan Frenkel (2004). "Designing specificity of protein-substrate interactions". In: *Phys. Rev. E* 70.5, p. 51917. DOI: 10.1103/PhysRevE.70.051917. URL: http://link.aps.org/doi/10.1103/PhysRevE.70.051917.

— (2007). "Monte Carlo Study of Substrate-Induced Folding and Refolding of Lattice Proteins". In: *Biophysical Journal* 92.4, pp. 1150–1156. ISSN: 0006-3495. DOI: http://dx.doi.org/10.1529/biophysj.106.084236. URL: http://www.sciencedirect.com/science/article/pii/S0006349507709268.

Cooper, J B et al. (1990). "X-ray analyses of aspartic proteinases". In: *Journal of Molecular Biology* 214.1, pp. 199–222. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/0022-2836(90)90156-G. URL: http://www.sciencedirect.com/science/article/pii/002228369090156G.

Daniel V. Schroeder (2000). *An introduction to Thermal Physics*. Robin J. Heyder, p. 159. ISBN: 0-321-27779-1.

Davtyan, Aram et al. (2012). "AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing". In: *The Journal of Physical Chemistry B* 116.29, pp. 8494–8503. DOI: 10.1021/jp212541y. URL: http://dx.doi.org/10.1021/jp212541y.

DeLano, Warren L (2002). "The PyMOL molecular graphics system". In:

Dias, Cristiano L et al. (2008). "Microscopic Mechanism for Cold Denaturation". In: *Phys. Rev. Lett.* 100.11, pp. 1–4. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.100.118101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.100.118101.

Dias, Cristiano L et al. (2010). "The hydrophobic effect and its role in cold denaturation". In: *Cryobiology* 60.1, pp. 91–99. ISSN: 0011-2240. DOI: 10.1016/j.cryobiol.2009.07.005. URL: http://www.sciencedirect.com/science/article/pii/S0011224009000996.

Dobson, Christopher M (2004). "Principles of protein folding, misfolding and aggregation". In: *Seminars in Cell & Developmental Biology* 15.1, pp. 3–16. ISSN: 10849521. DOI: 10.1016/j.semcdb.2003.12.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S1084952103001137.

Dragan, Anatoly I et al. (2004). "DNA binding and bending by HMG boxes: energetic determinants of specificity". In: *Journal of molecular biology* 343.2, pp. 371–393.

Durbin, S. D. and G. Feher (1996). "Protein Crystallization". en. In: *Annual Review of Physical Chemistry* 47.1, pp. 171–204. ISSN: 0066-426X. DOI: 10.1146/annurev.physchem.47.1.171. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.physchem.47.1.171.

Ehrnhoefer, Dagmar E et al. (2008). "EGCG redirects amyloidogenic polypeptides into unstructured, off-pathway oligomers". In: *Nat Struct Mol Biol* 15.6, pp. 558–566. ISSN: 1545-9993. URL: http://dx.doi.org/10.1038/nsmb.1437http://www.nature.com/nsmb/journal/v15/n6/suppinfo/nsmb.1437{_}S1.html.

Eisenhaber, Frank (1996). "Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures". In: *Protein Science* 5.8, pp. 1676–1686. ISSN: 1469-896X. DOI: 10.1002/pro.5560050821. URL: http://dx.doi.org/10.1002/pro.5560050821.

Faraggi, Eshel et al. (2012). "SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles." In: *Journal of computational chemistry* 33.3, pp. 259–67. ISSN: 1096-987X. DOI: 10.1002/jcc.21968. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3240697{\&}tool=pmcentrez{\&}rendertype=abstract.

Farber, Patrick et al. (2010). "Analyzing Protein Folding Cooperativity by Differential Scanning Calorimetry and NMR Spectroscopy". In: *Journal of the American Chemical Society* 132.17, pp. 6214–6222. ISSN: 0002-7863. DOI: 10.1021/ja100815a. URL: http://dx.doi.org/10.1021/ja100815a.

Floris, Matteo et al. (2011). "MAISTAS: a tool for automatic structural evaluation of alternative splicing products." In: *Bioinformatics* 27.12, pp. 1625–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr198.

Folch, Benjamin, Yves Dehouck, and Marianne Rooman (2010). "Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials." In: *Biophys. J.* 98.4, pp. 667–77. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2009.10.050. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2820637\&tool=pmcentrez\&rendertype=abstract.

Frock, AD and RM Kelly (2012). "Extreme thermophiles: moving beyond single-enzyme biocatalysis". In: *Curr. Opin. Chem. Eng.* 1.4, pp. 363–372. DOI: 10.1016/j.coche.2012.07.003.Extreme. URL: http://www.sciencedirect.com/science/article/pii/S2211339812000433.

Garg, Aarti, Harpreet Kaur, and G P S Raghava (2005). "Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure." In: *Proteins* 61.2, pp. 318–24. ISSN: 1097-0134. DOI: 10.1002/prot.20630. URL: http://www.ncbi.nlm.nih.gov/pubmed/16106377.

Gast, Klaus et al. (1994). "Compactness of protein molten globules: temperature-induced structural changes of the apomyoglobin folding intermediate". In: *European Biophysics Journal* 23.4, pp. 297–305. ISSN: 0175-7571. DOI: 10.1007/BF00213579. URL: http://dx.doi.org/10.1007/BF00213579.

Gemma Soldi, Francesco Bemporad, and and Fabrizio Chiti* (2008). "The Degree of Structural Protection at the Edge β-Strands Determines the Pathway of Amyloid Formation in Globular Proteins". In: *Journal of the American Chemical Society* 130.13, pp. 4295–4302. DOI: 10.1021/ja076628s. URL: http://dx.doi.org/10.1021/ja076628s.

Gómez, Javier et al. (1995). "The heat capacity of proteins". In: *Proteins: Structure, Function, and Bioinformatics* 22.4, pp. 404–412. ISSN: 1097-0134. DOI: 10.1002/prot.340220410. URL: http://dx.doi.org/10.1002/prot.340220410.

Gomez, Javier et al. (1995). "The heat capacity of proteins". In: *Proteins: Structure, Function, and Bioinformatics* 22.4, pp. 404–412.

Griep, Sven and Uwe Hobohm (2010). "PDBselect 1992–2009 and PDBfilter-select". In: *Nucleic Acids Research* 38.suppl 1, pp. D318–D319. DOI: 10.1093/nar/gkp786. URL: http://nar.oxfordjournals.org/content/38/suppl_1/D318.abstract.

Grossfield, Alan (2003). "WHAM: the weighted histogram analysis method".

Hallerbach, B and H.-J. Hinz (1999). "Protein heat capacity: inconsistencies in the current view of cold denaturation". In: *Biophysical Chemistry* 76.3, pp. 219–227. ISSN: 0301-4622. DOI: 10.1016/S0301-4622(98)00239-7. URL: http://www.sciencedirect.com/science/article/pii/S0301462298002397.

Halperin, Inbal et al. (2002). "Principles of docking: An overview of search algorithms and a guide to scoring functions." In: *Proteins* 47.4, pp. 409–43. ISSN: 1097-0134. DOI: 10.1002/prot.10115. URL: http://www.ncbi.nlm.nih.gov/pubmed/12001221.

Hazy, E et al. (2011). "Distinct Hydration Properties of Wild-Type and Familial Point Mutant A53T of α-Synuclein Associated with Parkinson's Disease". In: *Biophysical Journal* 101.9, pp. 2260–2266. DOI: 10.1016/j.bpj.2011.08.052. URL: http://www.cell.com/biophysj/abstract/S0006-3495(11)01061-7.

Hoang, Trinh Xuan et al. (2004). "Geometry and symmetry presculpt the free-energy landscape of proteins". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.21, pp. 7960–7964. DOI: 10.1073/pnas.0402525101. URL: http://www.pnas.org/content/101/21/7960.abstract.

Hobohm, U and C Sander (1994). "Enlarged representative set of protein structures". In: *Protein Sci* 3.3, pp. 522–524. URL: http://www.hubmed.org/display.cgi?uids=8019422.

Hobohm, Uwe et al. (1992). "Selection of representative protein data sets". In: *Protein Sci.* 1.3, pp. 409–417.

Huang, David M and David Chandler (2000). "Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding". In: *Proceedings of the National Academy of Sciences* 97.15, pp. 8324–8327. DOI: 10.1073/pnas.120176397. URL: http://www.pnas.org/content/97/15/8324.abstract.

Hubbard, T J and T L Blundell (1987). "Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling". In: *Protein Eng* 1.3, pp. 159–171. URL: http://www.hubmed.org/display.cgi?uids=3507702.

Hunter, John D (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 0090–95.

Ikenoue, Tatsuya et al. (2014). "Cold Denaturation of α-Synuclein Amyloid Fibrils". In: *Angewandte Chemie International Edition* 53.30, pp. 7799–7804. ISSN: 1521-3773. DOI: 10.1002/anie.201403815. URL: http://dx.doi.org/10.1002/anie.201403815.

Janin, Joel (1979). "Surface and inside volumes in globular proteins". In: *Nature* 277.5696, pp. 491–492. URL: http://dx.doi.org/10.1038/277491a0.

Jeppesen, Martin D et al. (2010). "A thermodynamic analysis of fibrillar polymorphism". In: *Biophysical Chemistry* 149.1–2, pp. 40–46. ISSN: 0301-4622. DOI: http://dx.doi.org/10.1016/j.bpc.2010.03.016. URL: http://www.sciencedirect.com/science/article/pii/S030146221000075X.

Johnson, Christopher M (2013). "Differential scanning calorimetry as a tool for protein folding and stability". In: *Archives of Biochemistry and Biophysics* 531.1–2, pp. 100–109. ISSN: 0003-9861. DOI: http://dx.doi.org/10.1016/j.abb.2012.09.008. URL: http://www.sciencedirect.com/science/article/pii/S0003986112003530.

Joo, Keehyoung, Sung Jong Lee, and Jooyoung Lee (2012). "Sann: solvent accessibility prediction of proteins by nearest neighbor method." In: *Proteins* 80.7, pp. 1791–7. ISSN: 1097-0134. DOI: 10.1002/prot.24074. URL: http://www.ncbi.nlm.nih.gov/pubmed/22434533.

Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22.12, pp. 2577–2637. ISSN: 1097-0282. DOI: 10.1002/bip.360221211. URL: http://dx.doi.org/10.1002/bip.360221211.

Kalia, Yogeshvar N et al. (1993). "The High-resolution Structure of the Peripheral Subunit-binding Domain of Dihydrolipoamide Acetyltransferase from the Pyruvate Dehydrogenase Multienzyme Complex of *Bacillus stearothermophilus*". In: *Journal of molecular biology* 230.1, pp. 323–341.

Kaliszan, Roman (1990). "High Performance Liquid Chromatographic Methods and Procedures of Hydrophobicity Determination". In: *Quantitative Structure-Activity Relationships* 9.2, pp. 83–87. ISSN: 09318771. DOI: 10.1002/qsar.19900090202. URL: http://doi.wiley.com/10.1002/qsar.19900090202.

Kardos, József et al. (2004). "Direct Measurement of the Thermodynamic Parameters of Amyloid Formation by Isothermal Titration Calorimetry". In: *Journal of Biological Chemistry* 279.53, pp. 55308–55314. DOI: 10.1074/jbc.M409677200. URL: http://www.jbc.org/content/279/53/55308.abstract.

Kato, Akio and Shuryo Nakai (1980). "Hydrophobicity determined by a fluorescence probe method and its correlation with surface properties of proteins". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 624.1, pp. 13–20. ISSN: 00052795. DOI: 10.1016/0005-2795(80)90220-2. URL: http://www.sciencedirect.com/science/article/pii/0005279580902202.

Kendrew, J C et al. (1958). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". In: *Nature* 181.4610, pp. 662–666. URL: http://dx.doi.org/10.1038/181662a0.

Kholodenko, Viktoriya and Ernesto Freire (1999). "A Simple Method to Measure the Absolute Heat Capacity of Proteins". In: *Analytical Biochemistry* 270.2, pp. 336–338. ISSN: 0003-2697. DOI: http://dx.doi.org/10.1006/abio.1999.4096. URL: http://www.sciencedirect.com/science/article/pii/S0003269799940964.

Kim, A and F C Szoka (1992). "Amino acid side-chain contributions to free energy of transfer of tripeptides from water to octanol." In: *Pharm. Res.* 9.4, pp. 504–14. ISSN: 0724-8741. DOI: 10.1023/A:1015892313856.

Kim, Aeri and FrancisC. C Szoka Jr (1992). "Amino Acid Side-Chain Contributions to Free Energy of Transfer of Tripeptides from Water to Octanol". In: *Pharm. Res.* 9.4, pp. 504–514. ISSN: 0724-8741. DOI: 10.1023/A:1015892313856. URL: http://dx.doi.org/10.1023/A:1015892313856.

Kim, Hai-Young et al. (2008). "Dissociation of Amyloid Fibrils of α-Synuclein in Supercooled Water". In: *Angewandte Chemie* 120.27, pp. 5124–5126. ISSN: 1521-3757. DOI: 10.1002/ange.200800342. URL: http://dx.doi.org/10.1002/ange.200800342.

Knowles, Tuomas P J et al. (2009). "An Analytical Solution to the Kinetics of Breakable Filament Assembly". In: *Science* 326.5959, pp. 1533–1537. ISSN: 0036-8075. DOI: 10.1126/science.1178250. URL: http://science.sciencemag.org/content/326/5959/1533.

Kuhn, M et al. (2014). "caret: classification and regression training. R package version 6.0-24". In: URL: https://scholar.google.nl/scholar?q=caret{\%}3A+classification+and+regression+training.+R+package+version+6.0-24{\&}btnG={\&}hl=nl{\&}as{_}sdt=0{\%}2C5{\#}0.

Kukic, Predrag et al. (2015). "Mapping the Protein Fold Universe Using the CamTube Force Field in Molecular Dynamics Simulations". In: *PLoS Comput Biol* 11.10, e1004435. DOI: 10.1371/journal.pcbi.1004435. URL: http://dx.doi.org/10.1371\%2Fjournal.pcbi.1004435.

Kyte, J and R F Doolittle (1982). "A simple method for displaying the hydropathic character of a protein." In: *J. Mol. Biol.* 157.1, pp. 105–32. ISSN: 0022-2836. URL: http://www.ncbi.nlm.nih.gov/pubmed/7108955.

Lomize, Mikhail A et al. (2006). "OPM: orientations of proteins in membranes database." en. In: *Bioinformatics (Oxford, England)* 22.5, pp. 623–5. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btk023. URL: http://bioinformatics.oxfordjournals.org/content/22/5/623.full.

Lum, Ka, David Chandler, and John D Weeks (1999). "Hydrophobicity at Small and Large Length Scales". In: *The Journal of Physical Chemistry B* 103.22, pp. 4570–4577. DOI: 10.1021/jp984327m. URL: http://pubs.acs.org/doi/abs/10.1021/jp984327m.

Marrink, Siewert J et al. (2007). "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations". In: *The Journal of Physical Chemistry B* 111.27, pp. 7812–7824. ISSN: 1520-6106. DOI: 10.1021/jp071097f. URL: http://dx.doi.org/10.1021/jp071097f.

May, Ali et al. (2014). "Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins". In: *Bioinformatics* 30.3, pp. 326–334. DOI: 10.1093/bioinformatics/btt675. URL: http://bioinformatics.oxfordjournals.org/content/30/3/326.abstract.

Meisl, Georg et al. (2016). "Molecular mechanisms of protein aggregation from global fitting of kinetic models". In: *Nat. Protocols* 11.2, pp. 252–272. ISSN: 1754-2189. URL: http://dx.doi.org/10.1038/nprot.2016.010http://10.0.4.14/nprot.2016.010http://www.nature.com/nprot/journal/v11/n2/abs/nprot.2016.010.html{\#}supplementary-information.

Mishra, Awadhesh K and Jagdish C Ahluwalia (1984). "Apparent molal volumes of amino acids, N-acetylamino acids, and peptides in aqueous solutions". In: *The Journal of Physical Chemistry* 88.1, pp. 86–92. DOI: 10.1021/j150645a021. URL: http://pubs.acs.org/doi/abs/10.1021/j150645a021.

Mitsutake, Ayori et al. (2004). "Combination of the Replica-Exchange Monte Carlo Method and the Reference Interaction Site Model Theory for Simulating a Peptide Molecule in Aqueous Solution". In: *The Journal of Physical Chemistry Bß* 108.49, pp. 19002–19012. DOI: 10.1021/jp047824d. URL: http://pubs.acs.org/doi/abs/10.1021/jp047824d.

Miyazawa, S and R L Jernigan (1985a). "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation". In: *Macromolecules* 18, pp. 534–552.

Miyazawa, Sanzo and Robert L Jernigan (1985b). "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation". In: *Macromolecules* 18.3, pp. 534–552. DOI: 10.1021/ma00145a039. URL: http://pubs.acs.org/doi/abs/10.1021/ma00145a039.

Monticelli, Luca et al. (2008). "The MARTINI Coarse-Grained Force Field: Extension to Proteins". In: *Journal of Chemical Theory and Computation* 4.5, pp. 819–834. ISSN: 1549-9618. DOI: 10.1021/ct700324x. URL: http://dx.doi.org/10.1021/ct700324x.

Morel, Bertrand, Lorena Varela, and Francisco Conejero-Lara (2010). "The Thermodynamic Stability of Amyloid Fibrils Studied by Differential Scanning Calorimetry". In: *The Journal of Physical Chemistry B* 114.11, pp. 4010–4019. DOI: 10.1021/jp9102993. URL: http://dx.doi.org/10.1021/jp9102993.

Morriss-Andrews, Alex and Joan-Emma Shea (2015). "Computational Studies of Protein Aggregation: Methods and Applications". In: *Annual Review of Physical Chemistry* 66.1, pp. 643–666. DOI: 10.1146/annurev-physchem-040513-103738. URL: http://dx.doi.org/10.1146/annurev-physchem-040513-103738.

Muñoz, Victor and Jose M Sanchez-Ruiz (2004). "Exploring protein-folding ensembles: A variable-barrier model for the analysis of equilibrium unfolding experiments". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.51, pp. 17646–17651. DOI: 10.1073/pnas.0405829101. URL: http://www.pnas.org/content/101/51/17646.abstract.

Murphy, Elizabeth C et al. (2001). "Structural basis for SRY-dependent 46-X, Y sex reversal: modulation of DNA bending by a naturally occurring point mutation". In: *Journal of molecular biology* 312.3, pp. 481–499.

Naganathan, Athi N. and Victor Muñoz (2014). "Thermodynamics of Downhill Folding: Multi-Probe Analysis of PDD, a Protein that Folds Over a Marginal Free Energy Barrier". In: *The Journal of Physical Chemistry B* 118.30. PMID: 24988372, pp. 8982–8994. DOI: 10.1021/jp504261g. eprint: http://dx.doi.org/10.1021/jp504261g. URL: http://dx.doi.org/10.1021/jp504261g.

Naganathan, Athi N et al. (2010). "Navigating the downhill protein folding regime via structural homologues". In: *Journal of the American Chemical Society* 132.32, pp. 11183–11190.

Naganathan, Athi N et al. (2011a). "Estimation of protein folding free energy barriers from calorimetric data by multi-model Bayesian analysis". In: *Phys. Chem. Chem. Phys.* 13.38, pp. 17064–17076. DOI: 10.1039/C1CP20156E. URL: http://dx.doi.org/10.1039/C1CP20156E.

Naganathan, Athi N et al. (2011b). "Estimation of protein folding free energy barriers from calorimetric data by multi-model Bayesian analysis". In: *Physical Chemistry Chemical Physics* 13.38, pp. 17064–17076.

Ni, Ran et al. (2013a). "Interplay between Folding and Assembly of Fibril-Forming Polypeptides". In: *Phys. Rev. Lett.* 111.5, p. 058101. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.111.058101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.111.058101.

Ni, Ran et al. (2013b). "Interplay between Folding and Assembly of Fibril-Forming Polypeptides". In: *Phys. Rev. Lett.* 111.5, p. 58101. DOI: 10.1103/PhysRevLett.111.058101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.111.058101.

Ni, Ran et al. (2015). "Competition between surface adsorption and folding of fibril-forming polypeptides". In: *Phys. Rev. E* 91.2, p. 22711. DOI: 10.1103/PhysRevE.91.022711. URL: http://link.aps.org/doi/10.1103/PhysRevE.91.022711.

Nieba, L et al. (1997). "Disrupting the hydrophobic patches at the antibody variable/constant domain interface: improved in vivo folding and physical characterization of an engineered scFv fragment." In: *Protein engineering* 10.4, pp. 435–44. ISSN: 0269-2139. DOI: 10.1093/protein/10.4.435. URL: http://www.ncbi.nlm.nih.gov/pubmed/9194169.

Oldfield, C J et al. (2005). "Comparing and combining predictors of mostly disordered proteins". In: *Biochemistry* 44.6, pp. 1989–2000. DOI: 10.1021/bi047993o. URL: http://www.hubmed.org/display.cgi?uids=15697224.

Patel, Bryan A., Pablo G. Debenedetti, and Frank H. Stillinger (2007). "Method for Efficient Computation of the Density of States in Water-Explicit Biopolymer Simulations on a Lattice†". In: *The Journal of Physical Chemistry A* 111.49, pp. 12651–12658. DOI: 10.1021/jp0761970. URL: http://dx.doi.org/10.1021/jp0761970.

Patel, Bryan A et al. (2007). "A Water-Explicit Lattice Model of Heat-, Cold-, and Pressure-Induced Protein Unfolding". In: *Biophysical Journal* 93.12, pp. 4116–4127. ISSN: 0006-3495. DOI: 10.1529/biophysj.107.108530. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2098741/.

— (2008). "The effect of sequence on the conformational stability of a model heteropolymer in explicit water". In: *The Journal of Chemical Physics* 128.17. DOI: http://dx.doi.org/10.1063/1.2909974. URL: http://scitation.aip.org/content/aip/journal/jcp/128/17/10.1063/1.2909974.

Periole, Xavier et al. (2012). "Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers". In: *Journal of the American Chemical*

Society 134.26, pp. 10959–10965. ISSN: 0002-7863. DOI: 10.1021/ja303286e. URL: http://dx.doi.org/10.1021/ja303286e.

Peter, M E (2010). "Targeting of mRNAs by multiple miRNAs: the next step." In: *Oncogene* 29.15, pp. 2161–4. ISSN: 1476-5594. DOI: 10.1038/onc.2010.59. URL: http://dx.doi.org/10.1038/onc.2010.59.

Prabhu, Ninad V and Kim A Sharp (2005). "Heat capacity in proteins". In: *Annu. Rev. Phys. Chem.* 56, pp. 521–548.

Privalov, P L et al. (1989). "Heat capacity and conformation of proteins in the denatured state". In: *Journal of Molecular Biology* 205.4, pp. 737–750. ISSN: 0022-2836. DOI: 10.1016/0022-2836(89)90318-5. URL: http://www.sciencedirect.com/science/article/pii/0022283689903185.

Privalov, Peter L and Anatoly I Dragan (2007). "Microcalorimetry of biological macromolecules". In: *Biophysical Chemistry* 126.1–3, pp. 16–24. ISSN: 0301-4622. DOI: 10.1016/j.bpc.2006.05.004. URL: http://www.sciencedirect.com/science/article/pii/S0301462206001608.

Privalov, Peter L and George I Makhatadze (1993). "Contribution of Hydration to Protein Folding Thermodynamics: II. The Entropy and Gibbs Energy of Hydration". In: *Journal of Molecular Biology* 232.2, pp. 660–679. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1006/jmbi.1993.1417. URL: http://www.sciencedirect.com/science/article/pii/S0022283683714178.

Privalov, Peter L et al. (1999). "The energetics of HMG box interactions with DNA: thermodynamics of the DNA binding of the HMG box from mouse sox-5". In: *Journal of molecular biology* 294.4, pp. 997–1013.

Pucci, Fabrizio and Marianne Rooman (2014). "Stability Curve Prediction of Homologous Proteins Using Temperature-Dependent Statistical Potentials". In: *PLoS Comput Biol* 10.7, e1003689. DOI: 10.1371/journal.pcbi.1003689. URL: http://dx.doi.org/10.1371\%2Fjournal.pcbi.1003689.

Rettich, Timothy R et al. (1981). "Solubility of gases in liquids. 13. High-precision determination of Henry's constants for methane and ethane in liquid water at 275 to 328 K". In: *J. Phys. Chem.* 85.22, pp. 3230–3237. DOI: 10.1021/j150622a006. URL: http://pubs.acs.org/doi/abs/10.1021/j150622a006.

Rezus, Y L A and H J Bakker (2007). "Observation of Immobilized Water Molecules around Hydrophobic Groups". In: *Phys. Rev. Lett.* 99.14, p. 148301. DOI: 10.1103/PhysRevLett.99.148301. URL: http://link.aps.org/doi/10.1103/PhysRevLett.99.148301.

Romero-Vargas Castrillon, Santiago et al. (2012). "Phase Behavior of a Lattice Hydrophobic Oligomer in Explicit Water". In: *The Journal of Physical Chemistry B* 116.31, pp. 9540–9548. DOI: 10.1021/jp3039237. URL: http://pubs.acs.org/doi/abs/10.1021/jp3039237.

Rose, G D et al. (1985). "Hydrophobicity of amino acid residues in globular proteins". In: *Science (80-.).* 229.4716, pp. 834–838. DOI: 10.1126/science.4023714. URL: http://www.sciencemag.org/content/229/4716/834.abstract.

Ross and Poirier (2004). "Protein aggregation and neurodegenerative disease". In: *Nat Med* 10 Suppl. S10–7.

Sali, Andrej, Eugene Shakhnovich, and Martin Karplus (1994). "Kinetics of Protein Folding : A Lattice Model Study of the Requirements for Folding to the Native State". In: *Journal of Molecular Biology* 235.5, pp. 1614–1638. ISSN: 0022-2836. DOI: DOI:10.1006/jmbi.1994.1110. URL: http://www.sciencedirect.com/science/article/B6WK7-45NSKPC-NB/2/4e65eab586d40b10589a2a57668c3f62.

Saric, Andela et al. (2016). "Physical determinants of the self-replication of protein fibrils". In: *Nat Phys* 12.9, pp. 874–880. ISSN: 1745-2473. URL: http://dx.doi.org/10.1038/nphys3828http://10.0.4.14/nphys3828http://www.nature.com/nphys/journal/v12/n9/abs/nphys3828.html{\#}supplementary-information.

Sasahara, Kenji, Hironobu Naiki, and Yuji Goto (2005). "Kinetically Controlled Thermal Response of β2-Microglobulin Amyloid Fibrils". In: *Journal of Molecular Biology* 352.3, pp. 700–711. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/j.jmb.2005.07.033. URL: http://www.sciencedirect.com/science/article/pii/S0022283605008211.

Scarsi, Marco, Nicolas Majeux, and Amedeo Caflisch (1999). "Hydrophobicity at the surface of proteins". In: *Proteins: Structure, Function and Genetics* 37.4, pp. 565–575. ISSN: 08873585. DOI: 10.1002/(SICI)1097-0134(19991201)37:4<565::AID-PROT7>3.0.CO;2-V. URL: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0134(19991201)37:4{\%}3C565::AID-PROT7{\%}3E3.0.CO;2-V/pdf.

Shakhnovich, E et al. (1991). "Protein folding bottlenecks: A lattice Monte Carlo simulation". In: *Phys. Rev. Lett.* 67.12, pp. 1665–1668. DOI: 10.1103/PhysRevLett.67.1665. URL: http://link.aps.org/doi/10.1103/PhysRevLett.67.1665.

Shakhnovich, E I (1994). "Proteins with selected sequences fold into unique native conformation". In: *Phys. Rev. Lett.* 72.24, pp. 3907–3910. DOI: 10.1103/PhysRevLett.72.3907. URL: http://link.aps.org/doi/10.1103/PhysRevLett.72.3907.

Shakhnovich, E I and A M Gutin (1993a). "A new approach to the design of stable proteins". In: *Protein Engineering* 6.8, pp. 793–800. DOI: 10.1093/protein/6.8.793. URL: http://peds.oxfordjournals.org/content/6/8/793.abstract.

— (1993b). "Engineering of stable and fast-folding sequences of model proteins". In: *Proceedings of the National Academy of Sciences* 90.15, pp. 7195–7199. URL: http://www.pnas.org/content/90/15/7195.abstract.

Shan, Bing et al. (2010). "The Cold Denatured State of the C-terminal Domain of Protein L9 Is Compact and Contains Both Native and Non-native Structure". In: *Journal of the American Chemical Society* 132.13, pp. 4669–4677. DOI: 10.1021/ja908104s. URL: http://pubs.acs.org/doi/abs/10.1021/ja908104s.

Shaytan, Alexey K, Konstantin V Shaitan, and Alexei R Khokhlov (2009). "Solvent accessible surface area of amino acid residues in globular proteins: correlation of apparent transfer free energies with experimental hydrophobicity scales." In: *Biomacromolecules* 10.5, pp. 1224–37. ISSN: 1526-4602. DOI: 10.1021/bm8015169. URL: http://www.ncbi.nlm.nih.gov/pubmed/19334678.

Shen, Min-yi and Andrej Sali (2006). "Statistical potential for assessment and prediction of protein structures". In: *Protein Sci.* 15, pp. 2507–2524. DOI: 10.1110/ps.062416606.Instead. URL: http://onlinelibrary.wiley.com/doi/10.1110/ps.062416606/full.

Shirota, Matsuyuki, Takashi Ishida, and Kengo Kinoshita (2009). "Analyses on hydrophobicity and attractiveness of all-atom distance-dependent potentials". In: *Protein Sci.* 18.9, pp. 1906–1915. ISSN: 1469-896X. DOI: 10.1002/pro.201. URL: http://dx.doi.org/10.1002/pro.201.

Sirovetz, Brian J, Nicholas P Schafer, and Peter G Wolynes (2015). "Water Mediated Interactions and the Protein Folding Phase Diagram in the Temperature–Pressure Plane". In: *The Journal of Physical Chemistry B* 119.34, pp. 11416–11427. DOI: 10.1021/acs.jpcb.5b03828. URL: http://dx.doi.org/10.1021/acs.jpcb.5b03828.

Smialowski, Pawel et al. (2007). "Protein solubility: sequence based prediction and experimental verification." In: *Bioinformatics* 23.19, pp. 2536–42. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl623. URL: http://www.ncbi.nlm.nih.gov/pubmed/17150993.

Soldi, Gemma et al. (2016). "Amyloid Formation of a Protein in the Absence of Initial Unfolding and Destabilization of the Native State". In: *Biophysical Journal* 89.6, pp. 4234–4244. ISSN: 0006-3495. DOI: 10.1529/biophysj.105.067538. URL: http://dx.doi.org/10.1529/biophysj.105.067538.

Spolar, R S, J H Ha, and M T Record (1989). "Hydrophobic effect in protein folding and other noncovalent processes involving proteins". In: *Proceedings of the National Academy of Sciences* 86.21, pp. 8382–8385. URL: http://www.pnas.org/content/86/21/8382.abstract.

Szyperski, Thomas et al. (2006). "Combined NMR-observation of cold denaturation in supercooled water and heat denaturation enables accurate measurement of ΔC p of protein unfolding". In: *European Biophysics Journal* 35, pp. 363–366. URL: http://dx.doi.org/10.1007/s00249-005-0028-4.

Tajima, Takahisa et al. (2013). "Construction of a simple biocatalyst using psychrophilic bacterial cells and its application for efficient 3-hydroxypropionaldehyde production from glycerol." In: *AMB Express* 3.1, p. 69. ISSN: 2191-0855. DOI: 10.1186/2191-0855-3-69. URL: http://www.ncbi.nlm.nih.gov/pubmed/24314120.

Thiriot, David S, Alexander A Nevzorov, and Stanley J Opella (2005). "Structural basis of the temperature transition of Pf1 bacteriophage". In: *Protein Sci.* 14.4, pp. 1064–1070. ISSN: 1469-896X. DOI: 10.1110/ps.041220305. URL: http://dx.doi.org/10.1110/ps.041220305.

Tran, Thanh Thuy, Phuong H Nguyen, and Philippe Derreumaux (2016). "Lattice model for amyloid peptides: OPEP force field parametrization and applications to the nucleus size of Alzheimer's peptides". In: *The Journal of Chemical Physics* 144.20. DOI: http://dx.doi.org/10.1063/1.4951739. URL: http://scitation.aip.org/content/aip/journal/jcp/144/20/10.1063/1.4951739.

Tusnády, Gábor E, Zsuzsanna Dosztányi, and István Simon (2005). "PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank." In: *Nucleic acids research* 33.Database issue, pp. D275–8. ISSN: 1362-4962. DOI: 10.1093/nar/gki002. URL: http://nar.oxfordjournals.org/content/33/suppl{_}1/D275.short.

Uversky, Vladimir N, Jie Li, and Anthony L Fink (2001). "Evidence for a Partially Folded Intermediate in α-Synuclein Fibril Formation". In: *Journal of Biological Chemistry* 276.14, pp. 10737–10744. DOI: 10.1074/jbc.M010907200. URL: http://www.jbc.org/content/276/14/10737.abstract.

Vajpai, Navratna et al. (2013). "High-pressure NMR reveals close similarity between cold and alcohol protein denaturation in ubiquitin". In: *Proc. Natl. Acad. Sci.* 110.5, E368–E376. DOI: 10.1073/pnas.1212222110. URL: http://www.pnas.org/content/110/5/E368.abstract.

Van Dijk, E et al. (2016a). "Cold denaturation of amyloid fibrils explained through the hydrophobic temperature dependence". In: *In preparation.*

Van Dijk, E et al. (2016b). "Heat Capacity Baseline Prediction Using BICEP". In: *In preparation.*

Van Dijk, Erik, Arlo Hoogeveen, and Sanne Abeln (2015). "The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures". In: *PLoS Comput Biol* 11.5, e1004277. DOI: 10.1371/journal.pcbi.1004277. URL: http://dx.doi.org/10.1371\%2Fjournal.pcbi.1004277.

Van Dijk, Erik et al. (2015). "Supplementary Material - Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation". In: *Physical review letters.*

— (2016c). "Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation". In: *Phys. Rev. Lett.* 116.7, p. 78101. DOI: 10.1103/PhysRevLett.116.078101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.116.078101.

Van Oss, C.J. (1995). "Hydrophobicity of biosurfaces — Origin, quantitative determination and interaction energies". In: *Colloids and Surfaces B* 5.3-4, pp. 91–110. ISSN: 09277765. DOI: 10.1016/0927-7765(95)01217-7. URL: http://www.sciencedirect.com/science/article/pii/0927776595012177.

Vangone, Anna and Alexandre M J J Bonvin (2015). "Contacts-based prediction of binding affinity in protein–protein complexes". In: *eLife* 4. Ed. by Michael Levitt, e07454. ISSN: 2050-084X. DOI: Vangone2015. URL: https://dx.doi.org/10.7554/eLife.07454.

Venselaar, Hanka et al. (2010). "Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces." In: *BMC Bioinformatics* 11.1, p. 548. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-548. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2992548\&tool=pmcentrez\&rendertype=abstract.

Wälti, Marielle Aulikki et al. (2016). "Atomic-resolution structure of a disease-relevant Aβ(1–42) amyloid fibril". In: *Proceedings of the National Academy of Sciences* 113.34, E4976–E4984. DOI: 10.1073/pnas.1600749113. URL: http://www.pnas.org/content/113/34/E4976.abstract.

Weeks, John D, David Chandler, and Hans C Andersen (1971). "Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids". In: *The Journal of Chemical Physics* 54.12.

Widom, B., P. Bhimalapuram, and Kenichiro Koga (2003a). "The hydrophobic effect". In: *Phys. Chem. Chem. Phys.* 5.15, p. 3085. ISSN: 1463-9076. DOI: 10.1039/b304038k. URL: http://xlink.rsc.org/?DOI=b304038khttp://dx.doi.org/10.1039/B304038K.

Widom, B, P Bhimalapuram, and Kenichiro Koga (2003b). "The hydrophobic effect". In: *Phys. Chem. Chem. Phys.* 5.15, pp. 3085–3093. DOI: 10.1039/B304038K. URL: http://dx.doi.org/10.1039/B304038K.

Wiggins, Philippa M. (1997). "Hydrophobic hydration, hydrophobic forces and protein folding". In: *Physica A: Statistical Mechanics and its Applications* 238.1-4, pp. 113–128. ISSN: 03784371. DOI: 10.1016/S0378-4371(96)00431-1. URL: http://www.sciencedirect.com/science/article/pii/S0378437196004311.

Wilkins, M. R. et al. (1998). "Two-dimensional gel electrophoresis for proteome projects: The effects of protein hydrophobicity and copy number". In: *Electrophoresis* 19.8-9, pp. 1501–1505. ISSN: 01730835. DOI: 10.1002/elps.1150190847. URL: http://doi.wiley.com/10.1002/elps.1150190847.

Wong, Kam-Bo, Stefan M V Freund, and Alan R Fersht (1996). "Cold Denaturation of Barstar:1H,15N and13C {NMR} Assignment and Characterisation of Residual Structure". In: *Journal of Molecular Biology* 259.4, pp. 805–818. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1006/jmbi.1996.0359. URL: http://www.sciencedirect.com/science/article/pii/S0022283696903599.

Wood, Graham R. et al. (2011). "Cotranslational protein folding and terminus hydrophobicity". In: *Advances in Bioinformatics* 2011. ISSN: 16878027. DOI: 10.1155/2011/176813.

Wright, P E and H J Dyson (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." In: *Journal of molecular biology* 293.2, pp. 321–331. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.3110. URL: http://www.sciencedirect.com/science/article/pii/S0022283699931108.

Wuttke, René et al. (2014). "Temperature-dependent solvation modulates the dimensions of disordered proteins". In: *Proceedings of the National Academy of Sciences* 111.14, pp. 5213–5218. DOI: 10.1073/pnas.1313006111. URL: http://www.pnas.org/content/111/14/5213.abstract.

Young, L, R L Jernigan, and D G Covell (1994). "A role for surface hydrophobicity in protein-protein recognition." In: *Protein science : a publication of the Protein Society* 3.5, pp. 717–29. ISSN: 0961-8368. DOI: 10.1002/pro.5560030501. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2142720{\&}tool=pmcentrez{\&}rendertype=abstract.

Zarrine-Afsar, Arash et al. (2008). "Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding". In: *Proceedings of the National Academy of Sciences* 105.29, pp. 9999–10004. DOI: 10.1073/pnas.0801874105. eprint: http://www.pnas.org/content/105/29/9999.full.pdf+html. URL: http://www.pnas.org/content/105/29/9999.abstract.

Zhou, Hongyi and Yaoqi Zhou (2002). "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction". In: *Protein Sci.* 11.11, pp. 2714–2726. ISSN: 1469-896X. DOI: 10.1110/ps.0217002. URL: http://dx.doi.org/10.1110/ps.0217002.

Zvelebil, M J and J O Baum (2008). *Understanding Bioinformatics*. Garland Science. ISBN: 9780815340249. URL: http://books.google.nl/books?id=dGayL_tdnBMC.

Summary of thesis

The central theme of this thesis is the temperature dependence of hydrophobicity in protein folding and aggregation. We have investigated the temperature dependence using three approaches. Firstly, we estimate the temperature dependence of the interactions using a theoretical and statistical approach (Chapter 2 and 3), and check if they are consistent with each other. Secondly, we introduce the temperature dependence of hydrophobic particles in a computational model (Chapter 3 and 6) and see what parameter ranges reproduce known experimental results. Thirdly, we verify that the estimates of the temperature dependence are consistent with the parameter ranges that reproduce experimental results, and use the conclusions to make predictions for real proteins (Chapter 4 and 5).

Below a short summary of each chapter.

In chapter 2 we use an approach that has been applied before to estimate pair wise amino acid interactions. This approach relies on the observation that, if two amino acids have a favourable interaction, they are more likely to be in contact in real structures. Since there are currently many protein structures available, we can also turn this around: If two amino acid types make a contact with each other more often than would be expected by random chance, they have a favourable interaction. This approach to derive a statistical potential was pioneered in (Miyazawa and Jernigan, 1985b).

Here, we have to make some adjustments: We focus on the interactions of amino acids and water, so we have to estimate the background probability of solvent accessibility of an amino acid type. Moreover, we are interested in temperature dependent effects. Therefor, we need to divide the structures in different temperatures. In this work, we only used NMR structures, since in X-ray structures the proteins are crystallised, and the entropic contributions of the hydrophobic effect are not taken into account properly.

We compare the results obtained in this fashion for hydrophobic amino acids with predictions and experimental observations for purely hydrophobic particles. We find that the interactions become weaker at lower and at higher temperatures, which is consistent with the theoretical predictions and experimental observations.

In the third chapter, we use the theoretical predictions in a coarse grained lattice model for protein folding. We find that incorporating the temperature dependence of the hydrophobic effect explains two additional phenomena when compared with a traditional pairwise potential. The first is cold denaturation of proteins. Most proteins denature at high temperatures, however some also denature at low temperatures. Heat induced denaturation can easily be explained by the higher chain entropy of the unfolded state. At higher temperatures, this entropic benefit becomes more important than the favourable entropy of the folded state.

The weaker hydrophobic interactions at lower temperatures can explain cold denaturation. Since the hydrophobic effect causes proteins to form a hydrophobic core and is the stabilising factor, decreasing these hydrophobic interactions

disturbs the stability of proteins at low temperatures. Adding this effect in our lattice model destabilises the protein at realistic temperatures.

The second phenomenon that can be explained by adding the temperature dependence to the hydrophobic interactions requires some additional explanation. We investigate the heat capacity of a protein. The heat capacity of a substance is defined as the amount of energy required to raise the temperature of the substance by one degree Kelvin. The heat capacity for most substances does not depend on the temperature. The change from ice to water, and from liquid water to water vapour, which are both examples of a phase transition, requires additional energy. This results in a peak in the heat capacity around the transition temperature. Such a peak can be found for almost all substances that undergo a phase transition.

The heat induced unfolding of a protein in a solution is no exception. The peak in the heat capacity is also found in simulations, both with and without the temperature dependent hydrophobic interactions. However, proteins in solution have the strange property that the heat capacity increases significantly with temperature even when the protein is not unfolding. This property has previously not been understood on a microscopic level. Adding the temperature dependent hydrophobic interactions allows us to reproduce this effect, suggesting that these interactions are indeed more realistic.

In chapter 4, we continue on the work of chapter 3. We use the observation that the hydrophobic interactions cause the increase in heat capacity, and take it one step further. If the increase in heat capacity is caused by the hydrophobic interactions, we should be able to predict the increase of the heat capacity of a protein if we know how much hydrophobic surface area is exposed to the surface. Using an existing dataset of proteins, we show that we can improve the predictions of a previously published method that has been used in the analysis of experimental data.

In chapter 5, we try to see if we can predict the hydrophobic surface area from a sequence. This should allow us to predict the heat capacity. While there exist tools that try to predict if a given residue is exposed to the surface, so far no tool has tried to predict the total hydrophobic surface area. Since the solvent exposure of amino acids depends on the solvent exposure of an other amino acid, this allows us to improve over a prediction that involves summing the predictions for individual amino acids.

In chapter 6, we use the "new interactions" in a lattice model for aggregation. This again allows us to reproduce cold denaturation of fibrils, something that has been found experimentally as well. We also do a parameter sensitivity study of several other parameters to investigate the thermodynamics of aggregation.

Nederlandse Samenvatting

Het centrale thema van deze thesis is de rol van de temperatuur afhankelijkheid van eiwitvouwing en aggregatie. We hebben deze temperatuur afhankelijkheid op drie manieren onderzocht. Eerst hebben we de grootte van de temperatuur afhankelijkheid geschat op twee manieren: een theoretische en een statistische aanpak (hoofdstuk twee en drie), en deze met elkaar vergeleken. Daarna hebben we deze schatting gebruikt in een computermodel en bekeken welke parameters de bekende experimentele resultaten reproduceren. De parameters die op deze manier geschat zijn hebben we gebruikt om een voorspelling te maken van de stabiliteit en warmtecapaciteit van echte eiwitten.

Hieronder volgt een korte samenvatting van elk hoofdstuk.

In hoofdstuk twee gebruiken we een bekende aanpak om de interactie tussen verschillende soorten aminozuren te schatten. Deze aanpak steunt op de observatie dat, als twee aminozuren een aantrekkende interactie hebben, ze vaker naast elkaar liggen in bestaande structuren. Omdat er veel verschillende bekende, vrij beschikbare, eiwitstructuren zijn, kunnen we deze observatie ook omdraaien: als twee soorten aminozuren vaker dan verwacht contact maken in bekende eiwitstructuren, hebben ze een aantrekkende interactie. Deze methode wordt toegepast op de structuren van eiwitten die bepaald zijn met verschillende temperaturen.

We hebben de interacties die zo berekend zijn voor de hydrofobe aminozuren vergeleken met theoretische en experimentele waarnemingen voor puur hydrofobe deeltjes. Uit deze vergelijking kan geconcludeerd worden dat de interacties van hydrofobe aminozuren zwakker worden op zowel lagere als hogere temperaturen. Deze conclusie is consistent met de experimentele en theoretische voorspellingen voor puur hydrofobe deeltjes.

In het derde hoofdstuk gebruiken we de theoretische voorspellingen in een grofmazig model voor eiwitvouwing. Het meenemen van de temperatuur afhankelijkheid van het hydrofobe effect verklaart twee verschijnselen die niet gevonden worden in het 'klassieke' model. De meeste eiwitten ontvouwen op hoge temperaturen, maar sommige ontvouwen ook op lage temperaturen. Ontvouwing op hoge temperaturen kan verklaard worden door de hogere entropie van de ongevouwen staat. Op hogere temperaturen wordt de entropie van de ongevouwen staat belangrijker dan de enthalpie van de gevouwen staat.

De zwakkere hydrofobe interacties op lagere temperaturen zijn de verklaring voor ontvouwing bij lage temperaturen. Op lage temperaturen zijn deze verzwakte hydrofobe interacties niet langer voldoende om een hydrofobe kern te vormen en het eiwit stabiel te houden. Dit wordt bevestigd in het roostermodel, waarbij het toevoegen van dit effect ervoor zorgt dat het eiwit op lage temperaturen inderdaad ontvouwt.

Het tweede fenomeen dat we kunnen verklaren met dit model vereist wat extra uitleg. We kijken naar de warmtecapaciteit van een eiwit. De warmtecapaciteit van een stof is gedefinieerd als de hoeveelheid warmte die nodig is om een stof met een graad Celcius te verhogen. De warmtecapaciteit van de meeste stoffen hangt niet af van de temperatuur. Met andere woorden, de energie die

nodig is om een stof van 20 naar 21 graden te verwarmen is hetzelfde als de benodigde energie om een stof van 80 naar een 81 graden te verwarmen. Een uitzondering hierop is een faseovergang, zoals bijvoorbeeld de verandering van ijs naar water. Een faseovergang kost extra energie. Dit veroorzaakt een piek in de warmtecapaciteit rond de transitie-temperatuur (0 graden Celsius voor water naar ijs). Een soortgelijke piek kan voor alle stoffen gevonden worden rond een transitie, zo ook het ontvouwen van een eiwit. De piek in de warmtecapaciteit wordt gevonden in zowel simulaties (met en zonder temperatuurafhankelijke hydrofobiciteit) als experimenten. Maar een eiwit, opgelost in water, heeft ook de vreemde eigenschap dat de warmtecapaciteit toeneemt met de temperatuur. Toevoeging van de temperatuurafhankelijkheid zorgt ervoor dat we dit effect kunnen reproduceren, wat suggereert dat deze temperatuurafhankelijkheid inderdaad de experimentele waarnemingen kan verklaren.

In hoofdstuk 4 wordt het werk van hoofdstuk 3 voortgezet. We gaan uit van de conclusie dat de stijging van de warmtecapaciteit veroorzaakt wordt door het hydrofobe effect. Dat suggereert dat het mogelijk moet zijn om de stijging van de warmtecapaciteit te voorspellen aan de hand van het hydrofobe oppervlak dat in contact is met water. We gebruiken een publieke dataset van eiwitten met hun warmtecapaciteit om te laten zien dat dit model nauwkeuriger is dan een eerder gepubliceerd model.

In hoofdstuk vijf ontwikkelen we een methode om het hydrofobe oppervlak van een eiwit in contact met water te voorspellen aan de hand van de samenstelling van het eiwit. Er bestaan al programma's die dit doen per aminozuur, maar in ons onderzoek blijkt het makkelijker om de totale hoeveelheid hydrofobe oppervlak van het eiwit te schatten. Dit oppervlak kan met de in hoofdstuk drie gevonden relatie gebruikt worden om de warmtecapaciteit te berekenen.

In hoofdstuk zes gebruiken we de temperatuurafhankelijke interacties, ontwikkeld in hoofdstuk 2 en 3, in een rooster model voor aggregatie. Dit zorgt ervoor dat het kapot gaan van 'fibrils' op lage temperaturen (wat in experimenten gevonden is) gereproduceerd kan worden in het computermodel.

Dankwoord

Een PhD is een groot persoonlijk project, maar het is anders dan de meeste mensen die niet in de wetenschap zitten zich voorstellen: je zit niet vier jaar in een kantoor opgesloten tot er een boekje uit komt, maar je werkt veel samen met verschillende mensen, binnen en buiten je eigen universiteit. Mijn thesis was niet mogelijk zonder de hulp van de volgende mensen:

Allereerst wil ik mijn begeleider, Sanne Abeln, bedanken. Sanne, het was een genoegen om met je te werken. Vanaf het allereerste begin van het project, mijn masterstage in Cambridge, via het VENI-voorstel waar mijn PhD op gebaseerd is, tot het schrijven van de artikelen, ben je nauw bij mijn werk betrokken geweest. Zonder jouw optimisme, enthousiasme en inzichten had dit boekje er heel anders uit gezien en was het er wellicht nooit geweest. Wat deze thesis naar mijn mening speciaal maakt, de diverse invalshoeken op hetzelfde probleem (Simulatie, Machine Learning, Data Mining en de "klassiek"-natuurkundige aanpak), was alleen mogelijk door jouw brede achtergrond en expertise op al deze gebieden.

Dat brengt me gelijk bij de tweede persoon die ik wil bedanken, Tuomas Knowles. In het begin van mijn masterproject waren we nog een beetje aan het zoeken naar een wetenschappelijke vraag. Ik legde uit waarom een eiwit ontvouwt in ons model op hoge temperaturen, en dat in ons model een eiwit nooit op lage temperaturen ontvouwt. Tuomas zijn reactie was: "Hoezo niet? Echte eiwitten doen dat wel. Zou het niet cool zijn als we dat in het model zouden kunnen vangen?". We liepen gelijk door naar het kantoor van Tuomas, waar nog een boek lag over eiwitvouwing. Het bleek dat ontvouwing op lage temperaturen direct gerelateerd is aan de richting waar mijn onderzoek sowieso naar op weg was: de temperatuurafhankelijkheid van het hydrofobe effect.

Ik wil ook mijn promotor, Jaap Heringa bedanken. Jaap is degene die me enthousiast heeft gemaakt over Bioinformatica tijdens een vak in mijn derde jaar. Zijn colleges maken complexe materie simpel, zonder de nuances te verliezen. Als ik aan Jaap denk moet ik aan een van zijn colleges denken. Het ging over het eiwit myoglobine, dat zuurstof transporteert van de longen naar de spieren. Jaap introduceerde het door het eiwit in een koe en een paard te laten zien, en te vragen: Welke deel van het eiwit denk je dat belangrijk is voor de functie? Het antwoord was dat het gedeelte van het eiwit dat hetzelfde was in een koe en een paard, cruciaal was voor de functie van het eiwit. Deze aanpak om de functie van een eiwit te bepalen was voor mij volledig nieuw, en de elegantie van het concept zorgde ervoor dat ik er zo snel mogelijk meer over wilde leren.

Verder wil ik mijn studenten, Maurits Dijkstra, Arlo Hoogeveen en Robbin Bouwmeester bedanken. Als jullie begeleider heb ik veel van jullie geleerd. Maurits, de enige persoon die twee keer in dit dankwoord voorkomt: Bedankt dat je mijn eerste student was. Arlo: jouw programmeerwerk en inzet hebben de basis gelegd voor Hoofdstuk 3. Tot slot, Robbin, bedankt voor je werk aan BICEP en HSApred, zowel binnen als buiten je stageperiode.

Ik wil ook graag de personen bedanken waarmee ik samengewerkt heb in Cambridge: Alexander, Patrick en Daan Frenkel. Alex, jouw inzicht in experimenten over eiwitaggregatie was enorm waardevol. Patrick, ik ken niemand anders die zoveel over het hydrophobic effect weet als jij. Daan, ik heb een tijd bij jouw group meetings gezeten en je bleef me verbazen. Ik heb nog nooit iemand ontmoet die consequent, bij elk project, een vraag kon stellen die de kern van het probleem liet zien.

Verder wil ik al mijn collega's bedanken, zowel binnen als buiten mijn afdeling op de VU. Annika, die me geholpen heeft bij de discussie. Maurits, met wie ik vaak nagepraat heb over het werk tijdens etentjes bij de Wagamama. Hierbij konden we even klagen over reviewers, om de volgende dag weer met nieuwe energie aan het werk te gaan. Verder wil ik al mijn andere collega's bedanken: Ali en Anton, met wie ik buiten mijn thesis gewerkt heb, en ook Punto, Bas, Hannes, Ted, Nicola en Qingzhen.

Buiten mijn collega's wil ik graag mijn vrienden bedanken: Chuck, Tim, Chris, Jolanda, Michel, Nigel, Jeroen, Sheung en Victor. Je ziet niet vaak dat een vriendengroep uit de middelbare school zo lang bij elkaar blijft. Ook wil ik alle tennissers van In Den Boogaerd waar ik mee getraind en gespeeld heb tijdens mijn PhD (in de voor- en najaarscompetitie) bedanken. De tenniscompetitie was altijd een mooie afwisseling op het werk dat ik doordeweeks deed.

Als laatste wil ik mijn familie bedanken. In het bijzonder mijn broer en zus, Leon en Ellen, en mijn ouders, Frank en Els. De vakanties, gesprekken aan de eettafel en humor die we met zijn allen delen is niet iets dat elk gezin heeft. Ik weet dat ik altijd op jullie kan steunen, in goede en slechte tijden.